universidad de león
Facultad de Ciencias Económicas y Empresariales

Facultad de Ciencias Económicas y Empresariales
Universidad de León

Grado en Marketing e investigación de mercados

Curso 2017/2018

ANÁLISIS DE SENTIMIENTO SOBRE LA CAPITALIDAD GASTRONÓMICA DE LA CIUDAD DE LEÓN A TRAVÉS DE TWITTER

SENTIMENT ANALYSIS ABOUT GASTRONOMIC CAPITAL FROM THE CITY OF LEÓN THROUGH TWITTER

Realizado por el alumno: D. Verónica Núñez García

Tutelado por los Profesores: Dr. D. Enrique López González
Dra. Dña. María Cristina Mendaña Cuervo

León, a 10 de julio de 2018

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE GRÁFICOS

ÍNDICE DE TABLAS

ÍNDICE DE CUADROS

RESUMEN

Con este trabajo se pretende obtener evidencias sobre el sentimiento de la capitalidad gastronómica de la ciudad de León. Para ello, se realizará un análisis de sentimiento que utiliza la inteligencia artificial y el aprendizaje automático. Se trata de un estudio que combina marketing y tecnología para dar a conocer una herramienta muy útil que puede suponer para las empresas una ventaja competitiva.

Las estrategias de social media no solo tienen que ver con promocionar los productos o servicios en las redes sociales, lo verdaderamente útil es extraer información de estas y convertirla en conocimientos de gran valor para la empresa.

Se utilizarán los tweets referentes a la capitalidad gastronómica de León para determinar cuál es el sentimiento por parte del público sobre este título, utilizando el software Rstudio, de manera que sea el propio ordenador el que tenga el poder de predicción suficiente para que a la hora de incorporar nuevos tweets sea éste el que determine el sentimiento en cuestión de segundos. Además, se repetirá el proceso para la capitalidad gastronómica de Huelva a efectos comparativos.

Finalmente, basado en las evidencias obtenidas, se extraen una serie de conclusiones referentes al objetivo principal del estudio y a los demás temas propiamente característicos de este trabajo.

PALABRAS CLAVE

Análisis de sentimiento, inteligencia artificial, inteligencia de negocios, marketing inteligente, big data, ciencia de datos, capitalidad gastronómica, máquina de vectores de soporte, algoritmo, aprendizaje automático, redes sociales.

ABSTRACT

This work aims to obtain evidence on the feeling of the gastronomic capital of the city of León. For this, a sentiment analysis using artificial intelligence and machine learning will be carried out. This research combines marketing and technology to make know a very useful tool which can provide a competitive advantage for companies.

Social media strategies do not only have to do with promoting products or services in social networks, what is useful is to extract information from these and turn it into knowledge of great value for the company.

The tweets referring to the gastronomy capital of León will be using to determine what the public´s feeling about this title. The software using is Rstudio, so that it is the computer itself that has enough predictive power to predict new tweets in a few seconds. In addition, the same process will be followed to determine the feeling of the gastronomic capital of Huelva for comparative purposes.

Finally, conclusions are drawn concerning the main objective of the research and the other issues that are characteristic for this work.

KEYWORDS

Sentiment analysis, artificial intelligence, business intelligence, marketing intelligence, big data, data science, gastronomic capital, support vector machine, algorithm, machine learning, social networks.

INTRODUCCIÓN

En la actualidad, la tecnología es un parte más de la sociedad, está presente en casi todos los aspectos de la vida cotidiana. Se podría decir que las personas no seríamos capaces de sobrevivir sin ella. Sin duda es una herramienta que facilita la vida de las personas. En marketing, es un gran aliado, ya que es un recurso fundamental para el éxito de la empresa u organización.

La forma de comunicación entre las personas ha cambiado sustancialmente con la aparición de la tecnología y las redes sociales. A través de ellas, las personas interactúan y comentan sobre multitud de temas de una manera espontánea y libre. Por tanto, se genera gran volumen de información en la red que tiene un fuerte valor para las empresas y las instituciones. Esta información es cambiante y muy abundante por lo que se necesitan herramientas que permitan gestionar todos estos datos para convertirlos en información valiosísima para las empresas.

Bajo este contexto, nadie podría ser capaz de analizar todos estos datos extrayendo conclusiones determinantes a tiempo para la empresa, por ello el uso de la tecnología es clave. "La inteligencia artificial es la simulación de procesos de inteligencia humana por parte de las máquinas. Estos procesos incluyen el aprendizaje (la adquisición de información y reglas para el uso de la información), el razonamiento (usando las reglas para llegar a conclusiones aproximadas o definitivas) y la autocorrección" (Rouse, 2017).

Gracias al uso de la inteligencia artificial es posible optimizar las estrategias de marketing para extraer el máximo rendimiento posible. Es posible, entre otras cosas, conocer el sentimiento de las personas sobre diversos temas (campañas de marketing, productos, servicios, etc.) de una manera objetiva y con un coste extremadamente bajo en comparación con la realización de encuestas.

La capitalidad gastronómica es un título que otorga la oportunidad de tener un gran beneficio para la ciudad que lo posee, ya que implica que el número de turistas aumente y, por tanto, que la ciudad tenga un mayor reconocimiento por su gastronomía. En el presente año 2018, León tiene a su disposición este título, por lo cual es de gran interés averiguar cuál es el sentimiento por parte del público acerca de León y su gastronomía utilizando una herramienta de inteligencia artificial que sea capaz de predecir el sentimiento de forma automática.

Por lo tanto, se van a analizar los comentarios realizados a través de la red social Twitter referentes a la capitalidad gastronómica de León durante el primer semestre del año 2018 con el objetivo final de establecer unos resultados, basados en las evidencias, que muestren cuál es el sentimiento mayoritario del público y qué aspectos son los más relevantes tanto en lo positivo como en lo negativo. Además, a efectos comparativos se establecerán las diferencias con la ciudad que disponía este título durante el mismo periodo del año pasado, Huelva.

Desarrollo del trabajo

El presente trabajo ha sido estructurado de la siguiente manera. Tras establecer una serie de objetivos y marcada la metodología a seguir se llevará a cabo el estudio que comienza con una parte teórica donde queda reflejada la importancia y actualidad del marketing basado en los datos, así como la explicación teórica del análisis de sentimiento. Seguidamente comienza la parte práctica donde se ha realizado dicho análisis utilizando la inteligencia artificial y el aprendizaje automático. Finalmente, se ha llegado a una serie de conclusiones, basadas en las evidencias obtenidas, que tratan de determinar cuál es el sentimiento acerca de la capitalidad gastronómica de León.

OBJETO DEL TRABAJO

El objetivo principal que se persigue con el presente Trabajo de Fin de Grado es determinar el sentimiento por parte del público a través de Twitter sobre la capitalidad gastronómica de León. Además, se persiguen una serie de objetivos que dependen del principal:

- Conocer la importancia de hacer marketing basado en los datos (data drive marketing).
- Determinar el potencial que tiene la inteligencia artificial y el aprendizaje automático para las estrategias de marketing.
- Trabajar y manejar, a través de la informática, gran cantidad de datos convirtiendo los mismos en información relevante y valiosa.
- Obtener las palabras más características de los tweets que mejor definan el sentimiento de la capitalidad gastronómica. A su vez, también se pretende determinar aquellas palabras que caracterizan a los tweets positivos y aquellas otras que caracterizan a los tweets negativos.

- Conocer los lugares desde los cuales se han enviado los tweets para determinar si la capitalidad tiene repercusión fuera de León.
- Representación gráfica y visual de todos los aspectos posibles del análisis.
- Establecer una comparación del sentimiento sobre la capitalidad gastronómica de León y de Huelva.
- Conocer la capacidad de Rstudio para la realización de estudios de mercados utilizando inteligencia artificial.

METODOLOGÍA

Para la elaboración del presente TFG se parte de un estado del arte basado en la necesidad de realizar un análisis o estudio sobre el título de capitalidad gastronómica León 2018 y su repercusión o impacto en la sociedad. Con el fin de realizar un estudio original y que me permitiera iniciarme en el mundo de la inteligencia artificial aplicada a los negocios comencé a seguir las siguientes pautas para realizar un análisis de sentimiento utilizando el software R y su conjunto de herramientas integradas que componen Rstudio.

Extracción de la información

Se trata de un campo de la ciencia del que no tenía grandes conocimientos a priori, por ello, antes de comenzar a realizar el estudio, ha tenido lugar esta primera y larga fase de investigación que se comprende desde entender cómo funciona Rstudio a descubrir cómo se usa la inteligencia artificial y cuáles son sus cimientos, sobre todo en lo referente al análisis de sentimiento utilizando la máquina de vectores de soporte (support vector machine). Por lo tanto, se ha obtenido gran cantidad de información haciendo referencia a fuentes secundarias tales como libros, revistas, tesis, cursos online, etc.

Parte teórica

Tras documentarme sobre todo lo referente a la elaboración del trabajo, se comienza a poner en evidencia todos los aspectos teóricos de la importancia de hacer marketing utilizando los datos y la oportunidad de conseguirlos de una manera relativamente fácil utilizando la información de las redes sociales. Además, la explicación teórica del análisis de sentimiento y sus grandes ventajas para el mundo del marketing y la investigación de mercados contribuye a dar más fuerza a la argumentación de la necesidad del uso y aprovechamiento de la gran cantidad de información que nada por la red.

Parte práctica

Es aquí donde comienzo a poner en práctica la mayoría de los conocimientos adquiridos en la primera fase de preparación. La extracción de los datos o scraping para la generación del modelo de clasificación es a través de una fuente primaria, a partir de Twitter Archiver se recogen todos los tweets que cumplan con las especificaciones de la búsqueda durante el primer semestre. A partir de ahí, se prepara la base de datos eliminando el ruido y clasificando los tweets como positivos o negativos. Se utiliza Rstudio para comenzar con el preprocesamiento de texto, y seguir con el análisis de sentimiento. Se ha usado la máquina de vectores de soporte, utilizando el hiperplano de separación que sitúe los tweets en un lado u otro en función de su polaridad. Una vez que ha sido construido el algoritmo de clasificación se evalúa el modelo con nuevos tweets para determinar el poder predictivo.

Ya concluido el análisis, se repite el proceso para la capitalidad gastronómica anterior, Huelva, con el único objetivo de la comparación con León.

Conclusiones

Tras la finalización del análisis de sentimiento, obtenidas las evidencias, es el momento de extraer una serie de conclusiones que vienen determinadas por los datos numéricos que proporciona Rstudio, así como por las representaciones gráficas y visuales. Dichas conclusiones tienen que ver con la determinación del sentimiento del público en cuanto a la capitalidad gastronómica pero también con la importancia y valor del marketing basado en datos y la inteligencia artificial para la consecución del éxito empresarial.

Capítulo I: DATA DRIVE MARKETING

1.1 LA ERA DE LOS DATOS

El mundo actual está marcado por el entorno digital donde cada día se generan millones de datos y la accesibilidad a ellos es relativamente fácil. Podría decirse que los datos son el petróleo del Siglo XXI (Herencia, 2018).

Una buena recolección, gestión y análisis de la información permite establecer una conexión inteligente entre el consumidor y la marca.

1.2 DEFINICIÓN DEL MARKETING BASADO EN DATOS

El marketing basado en los datos es el proceso mediante el cual los profesionales del marketing recogen conocimientos y tendencias, que son el resultado de los análisis de datos procedentes del mercado o las empresas. Después, se traducen esos conocimientos en decisiones para la empresa.

El principal objetivo del marketing basado en los datos es la optimización de las estrategias para entender las tendencias que cambian rápidamente, así como para satisfacer las demandas de los consumidores que son únicas. Se trata de aprovechar los datos para tener una visión mucho más profunda de lo que los clientes desean en cada momento (Stringfellow, 2018). En definitiva, un marketing sin una investigación de mercados previa a través de los datos que proporciona el mercado no tiene ninguna certeza de que dichos esfuerzos de marketing se van a convertir en un beneficio para la empresa u organización.

Cuando las marcas son capaces de entender perfectamente quién, qué, dónde, cuándo y por qué los consumidores o clientes interactúan con las decisiones comerciales conocen mejor a su mercado objetivo y están en condiciones óptimas de tomar mejores decisiones.

1.3 VENTAJAS DEL MARKETING BASADO EN DATOS

El marketing basado en los datos se fundamenta en el uso de la información que proporcionan los datos para mejorar las estrategias y decisiones. Así es más fácil invertir los recursos de marketing, maximizar el retorno de la inversión, reducir gastos, proporcionar al cliente un valor añadido, en definitiva, fidelizarle. Esta manera

de hacer marketing tiene muchas ventajas, y las principales son la efectividad de las estrategias comerciales y la facilidad de la implantación de las herramientas necesarias para el análisis de los datos (Stringfellow, 2018).

La información valiosa que se consigue gracias a los datos puede proporcionar ideas sobre cómo dirigirse al público objetivo, y, en consecuencia, conocerle mejor, además las marcas pueden saber con certeza qué usuarios se involucran más con los esfuerzos de marketing y a través de qué canales lo hacen. También, permite a la dirección conocer con exactitud qué productos o servicios funcionan mejor en el mercado y cuáles no, así se pueden solucionar problemas que puedan surgir de una manera más rápida y eficiente. Además, utilizar este marketing aumenta las ventas cruzadas y las ventas dirigidas. En definitiva, es una ventaja indiscutible a la hora de centrar los esfuerzos de marketing en el momento más oportuno.

En la actualidad, los datos impulsan la mayoría de las decisiones de marketing en el mundo altamente competitivo (Stringfellow, 2018). Se ha dejado de lado la intuición para el marketing, sustituyéndola por la tecnología avanzada y el aprendizaje automático para obtener la información valiosa a una mayor velocidad (Think with Google, 2016).

Se ha visto que son múltiples las ventajas que emergen del marketing basado en datos, pero a continuación se desarrollaran con mayor detalle tres de ellas.

1.3.1 Satisfacción del cliente

Gracias al conocimiento existente sobre los clientes, la inteligencia del marketing sabrá determinar el momento idóneo para ofrecerles información relevante y personalizada. Esta oferta, cuánto más personalizada mayor impacto tendrá en el cliente, por lo tanto, éste estará muy satisfecho (SAS, s. f.).

Además, reconocer con mayor exactitud las necesidades específicas de los clientes a través de múltiples canales dotará a la empresa de una buena reputación que se convertirá en nuevos clientes de mucho valor.

1.3.2 Costes del servicio y retorno de la inversión (ROI)

La analítica inteligente sabe detectar, de todas las comunicaciones que se generan para un único cliente, aquellas que más probabilidad tienen de prosperar y las que son más

rentables desde el punto de vista empresarial. Esta analítica es capaz de determinar aquello que tiene repercusión en los clientes y lo que no, además de conocer dónde se debería concretar el presupuesto para sacarle el máximo partido. Ello se traduce en que se ajustará el gasto del dinero en función del valor del cliente y la distribución del presupuesto de marketing será más eficiente, es decir, disminución de costes y maximización del ROI.

1.3.3 Aumento de las ventas cruzadas y dirigidas

En la actualidad, los clientes están saturados de mensajes con ofertas comerciales que en muchas ocasiones son agresivos e irrelevantes. A través de la analítica, se conoce con mayor exactitud el perfil comercial de los clientes y, por ello, las empresas pueden decidir qué productos son adecuados para ofrecerlos en venta cruzada y dirigida sin necesidad de molestarles.

1.4 ESTABLECER UN PERFIL DEL CLIENTE GRACIAS AL BIG DATA

Para hacer marketing a partir de los datos no se necesita un departamento de científicos especializados dotados de las herramientas más caras. Será el departamento de marketing quien se beneficie de la analítica de los datos. Una vez que se hayan instalado y configurado todas las herramientas necesarias para llevar a cabo esta labor, se tendrá una visión profunda de la situación de forma que los profesionales puedan entender y utilizar dichas herramientas. Al automatizar la obtención de la información se consigue un autoservicio y unos resultados muy visuales que colocan a la empresa en una situación de ventaja con respecto a sus competidores.

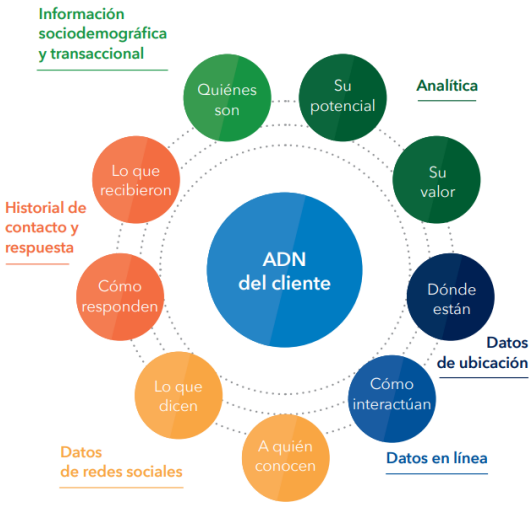
"Los cimientos de la analítica de marketing son los datos de clientes" (SAS, s. f., p. 12). Para triunfar no es necesario con recopilar el mayor número de información posible, hay que asegurarse de que la información que se está recopilando es la que interesa, es decir, los datos deben ser los adecuados, y a su vez, deben poder enriquecerse con otro tipo de datos.

A continuación, en la Figura 1.1 se mostrarán algunos ejemplos de los datos de clientes más comunes que se utilizan en el ámbito empresarial:

- Datos sociodemográficos.

- Historial de contacto y respuesta.
- Analítica.
- Datos de ubicación.
- Datos en línea.
- Datos de las redes sociales.

Figura 1.1 Información sobre el ADN del cliente



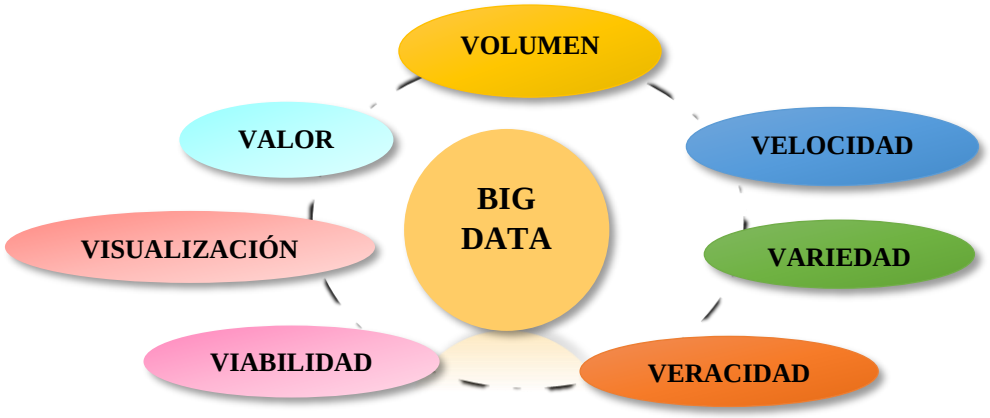
Fuente: (SAS, s. f., p. 12)

Una vez que se han conseguido los datos útiles, que corresponden a los datos explícitos, se debe utilizar una analítica para convertirlos en percepciones implícitas para que tengan sentido. Los datos explícitos son aquellos que están expresados directamente, por ejemplo, en un tweet: "¡El iPhone 7 no me gusta! ¡Es demasiado grande y se me cae al suelo todo el rato!". Ahora esta información se filtraría y se añadiría a los datos explícitos como son; producto: smartphone, problema: grande, sentimiento: negativo, marca: iPhone. Una vez que han sido analizados ofrecen las percepciones implícitas que son el sentimiento negativo y la tasa satisfacción de clientes negativa. De esta manera, la empresa se sumerge en el avanzado mundo del Big Data, caracterizado por su volumen masivo de datos en formatos diferentes que se multiplican y evolucionan a una velocidad increíble además de su análisis y actuación en consecuencia. Las V´s del Big data comenzaron siendo cuatro: volumen, velocidad, variedad y veracidad, pero lo cierto es que actualmente se puede hablar como mínimo de siete ya que se añaden tres más que son: la viabilidad, la visualización de los datos y el valor de estos (Instituto de ingeniería del conocimiento,

2016). A su vez, SAS también menciona la complejidad como una característica más del Big data (SAS, s. f.).

Por tanto, tal y como se muestra en la Figura 1.2, el Big data tiene múltiples características. Consta de gran cantidad de datos que se generan a cada segundo en el entorno (volumen), además estos datos están en constante movimiento, es decir, la rapidez con la que se crean, se almacenan y se procesan (velocidad). Estos datos están compuestos por diferentes formas tipos y fuentes, pueden ser estructurados o no estructurados (variedad), la incertidumbre de los datos trata de definir la fiabilidad de la información (veracidad), el uso eficiente del gran volumen de datos es muy importante para la inteligencia empresarial (viabilidad). Además, la complejidad de todo esto tiene que ser contrarrestada con una buena presentación de los datos (visualización), y, por último, la información obtenida de los datos que posteriormente se transformará en conocimiento es el verdadero valor de los datos y no éstos propiamente dichos (valor).

Figura1.2 Características del Big Data



Fuente: Elaboración Propia, basado en (Instituto de ingeniería del conocimiento, 2016) y (SAS, s. f.)

Es lógico, bajo este contexto, pensar que cuantos más datos relevantes se utilizan en una organización, cuantas más fuentes de diversos datos y cuanto más rápido se responda a ellos, más valiosa será la información que se produzca y las decisiones de marketing serán espectaculares. El grado de inteligencia y de analítica de mercado seguirá aumentando a medida que avanza la tecnología en lo referente a la ciencia de datos.

En la Figura 1.3 se ven las distintas aplicaciones que tiene la analítica de datos y sus diferentes etapas. Se ve como, en un primer lugar, lo que se realiza es un informe estándar detallando lo que ha ocurrido en un momento determinado. Después, es el momento de realizar los informes ad-hoc que son más específicos y tratan de determinar no solo lo que ha ocurrido sino la frecuencia y el lugar. Una vez realizados estos informes se generan las preguntas de investigación que buscan determinar con mayor exactitud cuál es el problema. Estas cuatro etapas corresponden a la fase de inteligencia de negocio. La analítica de negocios se ocupa de la parte estadística analizando lo que está ocurriendo, gracias a esas estadísticas se puede realizar una previsión de futuro para que la empresa pueda verse en el escenario de qué ocurrirá si las tendencias continúan. Además, establece una modelación predictiva determinando lo que ocurriría a continuación. Por último, realiza una optimización estableciendo qué sería lo mejor que podría pasar.

Figura1.3 Aplicaciones de la análitica de los datos.

Fuente: (SAS, s. f., p. 14)

La analítica tiene muchas aplicaciones donde es posible conseguir conocimiento e información privilegiada frente a la competencia en base al mercado objetivo. Alguna de ellas son las ventas dirigidas y cruzadas, la segmentación de clientes, predicción de indicadores (KPI), conocer el valor del ciclo de vida de los clientes, minería web, detección del fraude, optimización del marketing, análisis de ubicación, analítica de redes sociales, etc.

1.5 CAMPAÑAS DE MARKETING EN TIEMPO REAL

"El hecho de tener una visión de 360 grados de sus clientes, así como una visión más profunda de su comportamiento es solamente el primer paso hacia un enfoque de marketing más inteligente. El segundo paso consiste en convertir esa información tan valiosa en acciones centradas en el cliente." (SAS, s. f., p. 21).

Cuando una empresa realiza marketing en tiempo real está sacando el máximo rendimiento al poder del marketing. Lo que además busca es mantener una relación totalmente individualizada con los clientes.

En la actualidad, los clientes están rebosados de información que en muchas ocasiones no aporta nada a sus decisiones de compra. Es por ello tan relevante el marketing inteligente, porque se centra más en informar al cliente que en intentar venderle. Además, es muy importante que esa información sea en el momento adecuado y al cliente adecuado, de ahí su gran valor. El cliente busca un buen servicio de marketing, que le informen en tiempo real sobre aquello que satisfaga sus necesidades.

Las empresas que tienen posibilidad de éxito son aquellas que verdaderamente están centradas en el cliente, pero no solo el departamento de marketing tiene que ser el conjunto de la organización quien tenga esta visión, sino la empresa en su total conjunto. Por eso, todos los datos que se tengan sobre los clientes deben estar conectados, aunque provengan de departamentos que no sean de marketing. Gracias a ello se podrá conocer con exactitud al cliente para que incluso la empresa pueda adelantarse a sus necesidades. Para conseguir toda esa información se necesita una analítica a tiempo real.

1.5.1 Tiempo real

Tomar decisiones a tiempo real permite reaccionar a la empresa al momento y optimizar todas las interacciones con los clientes. Gracias al marketing a tiempo real las empresas pueden ser capaces de aportar información relevante a cada cliente a través de sus canales favoritos en el momento apropiado, acompañándole en todo momento del ciclo de vida del cliente. Este marketing, trata de cumplir con las expectativas del cliente.

La analítica en tiempo real, principalmente, toma sus bases en tres principios (SAS, s. f.).

1.5.1.1 Saber escuchar a los consumidores en tiempo real

Es el momento de la captura y procesamiento de los datos provenientes de los clientes a tiempo real. Para que una empresa sea capaz de reaccionar a tiempo real debe analizar los datos de un cliente y un contexto, en lugar de hacerlo con un conjunto de clientes a la vez. Es por ello por lo que se necesitan datos a tiempo real para que la empresa sea consciente en todo momento de las posibles necesidades que pudieran surgir en los clientes.

1.5.1.2 Contextualizar la información en tiempo real

Escuchar y ver lo que el cliente dice y hace en cada momento es muy importante, pero carece de sentido si la empresa no es capaz de añadirle un valor a esos datos. Los datos explícitos que pudieran resultar de un perfil de cliente, por ejemplo, deben transformarse de manera que se descarte el ruido y se puedan colocar esos datos en un contexto relevante de actuación. Este contexto puede ser muchas cosas: alguien pagando en una tienda en línea, haciendo operaciones en su cuenta bancaria, tuiteando a sus amigos, etc. Lo que se busca es combinar los datos en tiempo real con otras actuaciones del mismo cliente a través de otros canales, así se tendría un contexto completo del comportamiento de los clientes. De esta manera, la empresa conocerá mejor a sus clientes y sabrá cuáles son los productos o servicios más adecuados para satisfacer sus necesidades.

1.5.1.3 Tomar decisiones y reaccionar a tiempo real

Es en este punto donde todo el esfuerzo anterior cobra sentido para la empresa. Tomar una decisión acertada en un momento clave es sin duda una ventaja para la empresa. Por ejemplo, un cliente de una empresa X hace unas horas buscaba por internet desde su Tablet unas sandalias de verano, además la previsión meteorológica determina que se acerca un anticiclón que traerá temperaturas muy altas. Es el momento para que la empresa mande una promoción relevante a dicho cliente ofreciéndole calzado de verano.

1.6 TRABAJAR CON LOS DATOS

Una de las principales dificultades de trabajar con datos a tiempo real es la gestión de esa gran cantidad de datos que han de ser tratados con mucho cuidado y a su vez con mucha rapidez. El valor de estos datos no son los datos propiamente dichos, sino los acontecimientos que están recogidos en ellos.

Al trabajar con cantidades tan grandes de datos es muy importante la filtración de aquellos que no aportan una información valiosa y relevante para las decisiones de marketing. Por tanto, solamente se almacena la información importante, como por ejemplo las campañas de marketing para las que un cliente en concreto fue seleccionado.

1.7 LA ETICA EN EL MARKETING BASADO EN DATOS

Puede ser contraproducente para una empresa no actuar con tacto con sus clientes. El poseer datos a tiempo real de los clientes es una ventaja, pero si no se tiene prudencia puede resultar perjudicial para la marca.

Es imprescindible e importantísimo respetar la privacidad de los clientes y actuar con ética, sino puede ser que éstos piensen que le están espiando o vigilando. La empresa tiene que saber cuándo reaccionar y a qué, así como cuando es apropiado o no.

Capítulo II: ANÁLISIS DE SENTIMIENTO

2.1 CONTEXTUALIZACIÓN

Vivimos una realidad en la cual los medios digitales forman parte de nuestra vida cotidiana. Actualmente, no se conciben las relaciones sociales o profesionales sin tener ningún tipo de presencia en estos entornos online. Debido a la "explosión de la Web 2.0 muchos usuarios emplean los medios sociales para compartir sus opiniones y experiencias acerca de productos, servicios o personas" (Vilares, Alonso, y Gómez-Rodríguez, 2013, p. 2).

Bajo esta perspectiva, con este nuevo canal de comunicación, los consumidores y/o usuarios comparten sus sentimientos acerca de todo aquello que les interesa o preocupa. Además, en muchas ocasiones dejan constancia de su grado de satisfacción o insatisfacción sobre un determinado producto o servicio apelando, en algunos casos, a sus características más positivas o negativas. Consecuentemente, estos comentarios llegan a más personas que directa o indirectamente influye sobre sus pensamientos acerca de dicho bien o servicio.

Esta nueva realidad 2.0 es muy interesante y llamativa para el marco empresarial y/ institucional. Puede proporcionar multitud de ventajas si se usan las herramientas necesarias para conseguir información relevante a tiempo real. A su vez, también conlleva un mayor grado de implicación, por parte de las empresas, en cuanto a la transparencia de sus acciones y a la importancia de la opinión de sus clientes sobre sus prácticas comerciales. Además, gracias a estos medios sociales, la relación empresa-cliente (B2C, Business to Consumer) podría ser mucho más eficiente.

Hasta ahora, las empresas e instituciones, para medir el grado de satisfacción de los clientes sobre la adquisición de un producto o el trato en la prestación de un servicio, realizaban encuestas donde se medían unos determinados insights. Posteriormente, se hacían las encuestas a los clientes y se medían los resultados dando un grado de satisfacción.

2.1.1 Inconvenientes del uso de encuestas en la investigación de mercados

Esta práctica que aún se lleva a cabo en la realidad, tiene unos grandes inconvenientes:

- Para conseguir datos sobre una muestra representativa de un público específico se necesita llegar a un gran número de personas por lo que requiere gran cantidad de recursos económicos y tiempo.
- Los encuestados inconscientemente están sesgados porque previamente se les ha hecho una pregunta, es decir, no expresan sus sentimientos de una manera natural, sino que la encuesta le pregunta sobre ciertos aspectos que éstos deben contestar. Además, esta falta de sinceridad puede deberse a la tendencia a proteger la privacidad por parte de los encuestados.
- Los encuestados por falta de tiempo e interés, en ocasiones muestran respuestas poco concienzudas ya que es frecuente no acabar de leer la pregunta o contestar demasiado rápido sin pensar.
- La interpretación de las preguntas por parte de los encuestados no siempre es de forma correcta y similar para todos ellos. Es por ello por lo que interviene un alto índice de subjetividad. Además, en un cuestionario es más complicado registrar los sentimientos de los encuestados, sobre todo si el investigador no está cara a cara para comprobar la expresión facial o el lenguaje corporal.
- La falta de personalización de los cuestionarios motiva al encuestado, en muchas ocasiones, a aburrirse y abandonar. Además, están hartos de rellenar encuestas y puede considerarse un factor de insatisfacción.

Es por ello necesario dar un paso más en cuanto a investigación se refiere. Conocer el sentimiento de las personas sobre diversos temas es ahora más sencillo gracias a las redes sociales y, en consecuencia, su grado de satisfacción.

2.2 ¿QUÉ ES EL ANÁLISIS DE SENTIMIENTO?

El análisis de sentimiento también se conoce como minería de opinión (opinión minning), extracción de opinión, minería de sentimiento, análisis de la subjetividad o análisis de la emoción y tiene como principal objetivo averiguar cuál es el tono emocional que hay detrás de una serie de palabras, es decir, trata de definir herramientas automáticas que sean capaces de extraer la información subjetiva de textos en lenguaje natural (sentimientos, opiniones) con el fin de crear un conocimiento estructurado que pueda ser utilizado por un sistema capaz de tomar decisiones.

Además, ha sido una de las áreas de investigación más activas en lo que se refiere al procesamiento del lenguaje natural (PLN) desde principios del año 2000 (Pozzi, Fersini, Messina, y Bing, 2017). Se utiliza, generalmente, para entender las emociones y actitudes que expresan un comentario en vía online. Trata, por tanto, de comprender una información que es completamente cualitativa además de generar indicadores de opinión subjetivos de una manera muy frecuente y con un coste mucho más reducido (Martínez Gordillo, 2016).

Debido a la importancia del análisis de sentimiento en los negocios y en la sociedad, su desarrollo se ha extendido desde la informática a las ciencias sociales. Durante los últimos años, la actividad industrial que rodea al análisis de sentimiento ha prosperado, además, grandes empresas como IBM, Microsoft, Google y SAS Global Communications han incorporado esta tecnología a su actividad comercial.

El análisis de sentimiento ha cobrado aún más importancia gracias a las redes sociales. La interconexión tan densa que se genera con frecuencia entre los usuarios activos genera un debate que puede ser capaz de motivar e implicar a los individuos en objetivos comunes y facilitar diversas formas de acción colectiva. La masiva información que se genera en las redes sociales es muy rica para usar en el análisis de sentimiento. Estos datos son datos no estructurados pero conllevan una ventaja muy importante para la ciencia de datos (Martínez Gordillo, 2016), son , por ejemplo, los que provienen de Twitter tales como imágenes, emoticonos, grabaciones… Estos datos no estructurados son los que más abundan en la actualidad. La enorme cantidad de datos que se generan de forma continua en las redes sociales requiere cambios radicales, basados en la convergencia de un área multidisciplinar que engloba las ventajas de la psicología, la sociología, el procesamiento del lenguaje natural y de aprendizaje automático.

El conocimiento existente en el contenido de la red social es de gran importancia tanto desde el punto de vista del usuario como desde el punto de vista de la empresa u organización ya que los usuarios pueden hablar sin prejuicios ni restricciones sobre cualquier tipo de tema lo que para la empresa es una fuente de información primaria valiosísima.

Con el análisis de sentimiento se busca la clasificación de las opiniones en dos categorías fundamentales: positivo y negativo (Kharde y Sonawane, 2016), aunque también se podría considerar una tercera categoría como es la neutral.

La analítica de texto trata de convertir todos esos datos en un lenguaje que el ordenador sea capaz de comprender para su posterior análisis. En este contexto, son muy necesarios los ordenadores para realizar este trabajo ya que una persona sería incapaz de analizar los aproximadamente 500 millones de tweets que se generan al día y mostrar unas conclusiones determinantes en el momento adecuado.

2.2.1 Procesamiento del lenguaje natural (PLN)

Se define como el campo que combina las tecnologías de la ciencia computacional (inteligencia artificial, aprendizaje automático o la inferencia estadística) con la lingüística aplicada (A. Moreno, 2017).

Su principal objetivo es hacer posible la comprensión y el procesamiento, asistidos por el ordenador, de información que está expresada en lenguaje humano para tareas concretas, como pueden ser la traducción automática, los sistemas de dialogo interactivos, los análisis de sentimiento, etc.

En cuanto a su aplicación en el análisis de sentimiento, el procesamiento del lenguaje natural se ocupa del análisis computacional de los textos que son producidos por las personas con el objetivo de recopilar unos niveles cuantificables de opiniones y sentimientos. Es decir, queremos que los ordenadores entiendan nuestro lenguaje para que éstos lo puedan procesar y generar así una información extremadamente valiosa (Martínez Gordillo, 2016).

A continuación, se explicará brevemente algunos de los componentes del procesamiento del lenguaje natural (A. Moreno, 2017):

- Análisis morfológico o léxico: consististe en el análisis interno de las palabras que componen una frase tratando de extraer lemas, rasgos flexivos y unidades léxicas compuestas. Es muy importante para entender la información básica: categoría sintáctica y significado léxico.
- Análisis sintáctico: es el análisis de la estructura de las oraciones de acuerdo con el modelo gramatical empleado que puede ser lógico o estadístico.
- Análisis semántico: consiste en la interpretación de las oraciones, una vez que han sido eliminadas las ambigüedades morfosintácticas.

- Análisis pragmático: recoge el análisis del contexto de uso a la interpretación final, es decir, incluye el tratamiento del lenguaje figurado como son las metáforas y las ironías.

2.3 LENGUAJE USADO EN LAS REDES SOCIALES

El análisis de sentimiento a través de una red social lleva incorporadas una serie de complejidades, como son los mensajes cortos, el ruido (todo aquello que no aporta información) o los metadatos incorporados a cada texto como son el género, la ubicación y la edad.

Los desafíos diarios a los que se enfrenta el análisis de sentimiento tienen que ver con la constante evolución del lenguaje que el usuario utiliza en las redes sociales ya que se tiende a usar aquellas palabras que nos rodean a diario. La evolución de este lenguaje se debe, en parte, a que a diario se interactúa con la tecnología y la mayor parte del lenguaje escrito que los usuarios perciben lo hacen a través de pantallas tales como móviles, tablets u ordenadores (Pozzi et al., 2017). Además, el lenguaje que se utiliza en las redes sociales para la comunicación entre los usuarios es de carácter más informal y personal.

Bajo este marco, para que el análisis de sentimiento consiga los resultados esperados, los sistemas que lo componen deben adaptarse de una manera nativa a la evolución continua del mensaje. Si no fuera capaz por sí mismo, la adaptación está en manos de los investigadores. La evolución del lenguaje influye significantemente en el uso que los usuarios hacen de la ironía y el sarcasmo.

Otro reto diario al que hay que enfrentarse para realizar este tipo de análisis es que las redes sociales, por definición, son dinámicas y heterogéneas donde las entidades implicadas interactúan entre sí. Por otro lado, considerar una representación de los datos del mundo real donde los datos son considerados como homogéneos, independientes e idénticamente distribuidos conduce a la introducción de un sesgo estadístico (Pozzi et al., 2017). Es por ello de extrema importancia combinar el contenido del texto con las relaciones entre los usuarios para realizar un adecuado análisis de sentimiento.

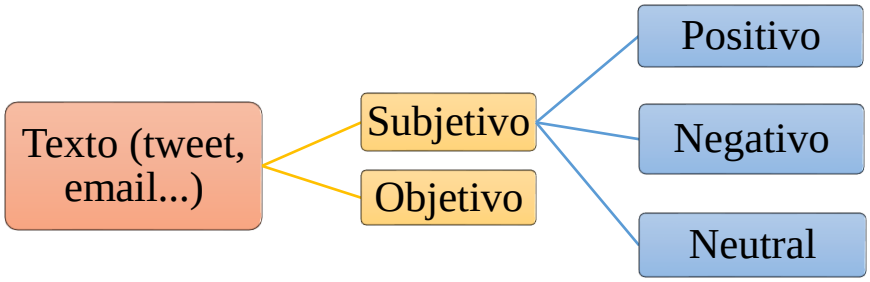
2.4 CARACTERÍSTICAS DEL ANÁLISIS DE SENTIMIENTO

El análisis de sentimiento pertenece a un amplio y complejo campo de investigación. Se desarrollarán, a continuación, con detalle las principales características.

2.4.1 Clasificación del sentimiento

El primer reto al que un investigador tiene que enfrentarse cuando va a realizar un análisis de sentimiento es determinar si el texto es de carácter objetivo o subjetivo. Como se ve en la Figura 2.1, cuando la oración es objetiva no se necesita realizar ninguna tarea más. Sin embargo, cuando ésta sea subjetiva, se necesita saber su polaridad, es decir, si es negativa, positiva o neutral.

Figura 2.1 Clasificación del sentimiento



Fuente: Elaboración propia

Se entienden por frases objetivas aquellas que expresan solamente una información sin dar ninguna opinión o sentimiento. Las frases subjetivas son aquellas que expresan una información con opiniones o sentimientos que pueden ser positivos, negativos o neutros. Un ejemplo de texto objetivo sería: “El iPhone 7 es un teléfono inteligente”, un ejemplo de texto subjetivo sería: “El iPhone 7 es asombroso”.

2.4.2 Los niveles del análisis

Ya se ha dicho que el propósito del análisis de sentimiento es la definición de herramientas automáticas capaces de entender la información subjetiva a partir del texto del lenguaje natural. El primer paso que hay que dar cuando se hace el análisis de sentimiento es comprender lo que quiere decir el texto que se está estudiando, el objeto de estudio. Por lo general, la información que se puede extraer de las redes sociales la podríamos dividir en tres niveles que se explicarán a continuación y a través de la figura 2.2.

Figura 2.2 Niveles del texto

Fuente: Elaboración propia basado en (Pozzi et al., 2017)

Como se ve en la figura anterior cada texto se puede desgranar en 3 niveles.

2.4.2.1 Nivel de mensaje

Lo que se hace en este nivel es determinar la polaridad, siempre que el texto sea subjetivo. Se analiza el texto en su conjunto y se determina si es positivo o negativo.

2.4.2.2 Nivel de oración

Es el momento de desgranar el texto por frases y determinar cuál es el sentimiento de cada una de ellas. Lo supuesto es que cada frase que compone un mensaje denota una sola opinión en una sola entidad (Pozzi et al., 2017).

2.4.2.3 Nivel de entidad y aspecto

Realiza un análisis más exhaustivo que el nivel del mensaje y el de oración. Se basa en la idea de que una opinión consiste en un sentimiento y una característica objetiva del producto o servicio. Por ejemplo: “El iPhone es impresionante, pero deberían trabajar más en el uso de la batería y en los problemas de privacidad”. Se evalúan 3 aspectos; el iPhone 7 es positivo, la batería es negativo y la privacidad también es negativo.

2.4.3 Opiniones regulares y opiniones comparativas

Las opiniones de los usuarios en las redes sociales u otros medios pueden ser de diferentes tonos que serán explicados a continuación.

2.4.3.1 Opinión regular

La opinión regular se refiere en la literatura a menudo como la opinión estándar y se divide en dos subtipos (Pozzi et al., 2017).

Por un lado, está la opinión directa que es aquella que es expresada de forma directa por una entidad, por ejemplo; "los gráficos de la pantalla del iPhone 7 son increíbles". Expresa su opinión de manera clara y literal.

Por otro lado, está la opinión indirecta que es aquella en la que una entidad expresa su opinión de forma indirecta sobre otras entidades, por ejemplo; "después de comprarme un iPhone he perdido todos mis datos de contacto". Este usuario describe un efecto no deseado, lo que provoca en él un sentimiento negativo.

2.4.3.2 Opinión comparativa

Una opinión es comparativa cuando expresa una relación de similitud o diferencia entre dos o más entidades, además, expresa una preferencia sobre una de ellas. Por ejemplo; "Apple es mucho más rápido que Samsung". "Una opinión comparativa se expresa generalmente con el uso de la forma comparativa o superlativa de un adjetivo o un adverbio" (Pozzi et al., 2017, p. 7).

2.4.4 Opiniones explícitas y opiniones implícitas

Una opinión puede asumir diferentes matices como son las opiniones implícitas o explicitas.

2.4.4.1 Opiniones explícitas

Se habla de una opinión explícita cuando la declaración es de carácter subjetivo, regular y/o comparativo. Por ejemplo; "la calidad de la cámara de fotos del iPhone 7 es impresionante".

2.7.2 Opiniones implícitas

Una opinión es implícita cuando la declaración es de carácter objetivo, regular y/o comparativo y que normalmente expresa un hecho determinado que es deseado o no por el autor. Por ejemplo; "el lunes por la mañana voy a ir a comprar el iPhone 7. No puedo

aguantar más tiempo sin él". Este comentario sugiere que hay buenas expectativas en cuanto al producto, aunque no se exprese con palabras.

2.4.5 Figuras literarias en las opiniones

Las figuras literarias, también conocidas como figuras retóricas son formas no convencionales de expresar opiniones dotadas con mayor vivacidad o belleza para sorprender o emocionar.

Las figuras más problemáticas de analizar en el procesamiento del lenguaje natural son la ironía y el sarcasmo. En la ironía, el autor da a entender una cosa cuando en realidad desea expresar justo lo contrario. Se usa, básicamente, para expresar giros del destino inesperados. El sarcasmo se utiliza, comúnmente, para transmitir una crítica implícita con un objetivo particular.

La dificultad de detectar el sarcasmo y la ironía provoca malentendidos en la comunicación entre las personas y plantea también problemas en los sistemas de procesamiento del lenguaje natural. Bajo la perspectiva del análisis de sentimiento, la ironía y el sarcasmo suelen considerarse como sinónimos. Cuando una frase contiene sarcasmo o ironía y el sistema de PLN lo considera como sentimiento positivo, suele ser en la realidad sentimiento negativo y viceversa.

2.5 LAS RELACIONES EN LAS REDES SOCIALES

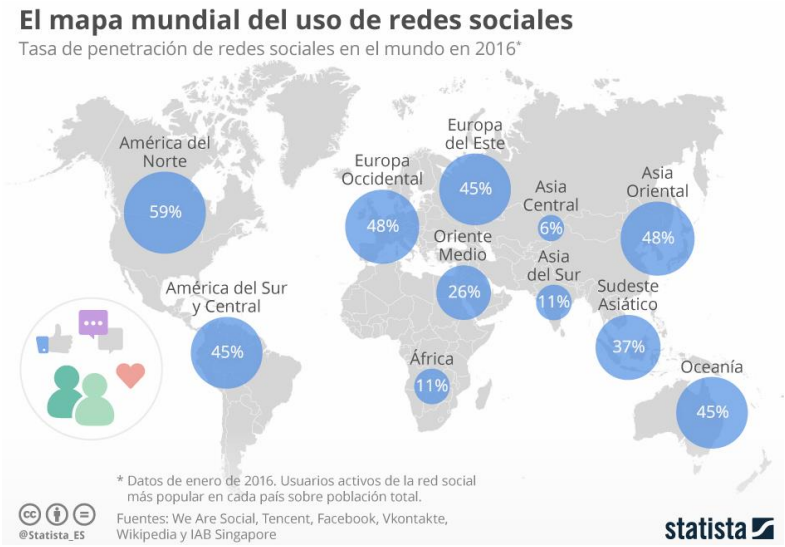
El análisis de sentimiento en las redes sociales tiene la premisa de que las opiniones que se generan por los diferentes usuarios son independientes e idénticamente distribuidas. Para averiguar realmente de donde procede la naturaleza del contenido de la red social hay que hacer referencia al principio de homofilia que se centra, básicamente, en que existe una tendencia entre los individuos de asociarse y relacionarse con aquellos que son semejantes a el mismo.

En este contexto, se puede entender que las relaciones de "amistad" entre los usuarios conllevan una tendencia de que tengan opiniones similares. Sin embargo, un sistema de análisis de sentimiento debe tener en cuenta que las hipótesis sobre las relaciones de amistad pueden no reflejar adecuadamente la realidad, donde los usuarios podrían tener opiniones diversas frente a un tema común (Pozzi et al., 2017).

2.5.1 Definición y uso de las redes sociales en línea

Puede considerarse como un nuevo canal de comunicación basado en las tecnologías digitales. En la actualidad, el uso de las redes sociales como medio de comunicación e interacción entre los usuarios aumenta a diario. Como se puede ver en el Gráfico 3.1, las redes sociales son usadas por todas las zonas más desarrolladas del planeta. Se ve como en América del Norte, por ejemplo, casi un 60% de la población utiliza este medio de comunicación digital, seguido por Europa Occidental con un 48%, Asia Oriental un 48%, América Central y Sur y Oceanía. Se puede concluir que, en 2016, aproximadamente la mitad de la población de los países desarrollados utiliza las redes sociales en su vida personal y/o profesional.

Gráfico 3.1 Mapa mundial del uso de las redes sociales en 2016



Fuente: (G. Moreno, 2016)

Existen 3 elementos que construyen una red social:

- La existencia de un espacio virtual en el que cada usuario pueda presentar su perfil personal y éste pueda ser accesible a los demás usuarios.
- La existencia de la posibilidad de contactar con el resto de los usuarios, es decir, de crear una red de contactos.
- La existencia de la posibilidad de análisis de las características de un perfil, en concreto, su red de contactos.

En definitiva, los espacios que forman las redes sociales se definen como los sitios webs donde los usuarios pueden crear un perfil público o semi público dentro de un sistema,

además, pueden tener una lista de contactos con los que tener comunicación, pudiendo además ver la lista de contactos de los demás dentro del sistema (Pozzi et al., 2017).

2.5.1.1 Tipos de relaciones entre los usuarios

Existen muchos tipos de redes sociales en línea y, por lo tanto, también las relaciones entre los usuarios son distintas. Las redes sociales pueden ser consideradas como espacios donde los usuarios pueden reunirse y comunicarse entre sí. El tipo de relación que los usuarios pueden tener en una red social se puede clasificar de la siguiente manera.

- Una relación bidireccional o de amistad: este tipo de relaciones son características de Facebook donde los usuarios son amigos entre sí y pueden ponerse en contacto a través de un chat privado, leer los mensajes del perfil de los demás amigos y conocer las actividades dentro de la red social. Este mecanismo permite el uso de una red social cerrada, es decir, nadie sin previo consentimiento puede acceder a la información personal. Además, ningún usuario podría considerarse como un desconocido ya que puede ser amigo de otra persona.
- Una relación denomina como "estrella": es el caso de Twitter donde se distinguen perfectamente el emisor del receptor. El emisor del mensaje puede ser el público en general, o bien dirigirse a un perfil en concreto. Con este modo de conexión un usuario puede ser tanto emisor como receptor de un mensaje. Podría considerarse como una red social abierta donde el mensaje se dirige desde un usuario a muchos otros usuarios que pueden tener cierto interés en el tema, aunque no compartan la misma opinión.

Capítulo III: PROYECTO LEÓN CAPITAL GASTRONOMICA 2018

3.1 RECOLECCIÓN DE DATOS

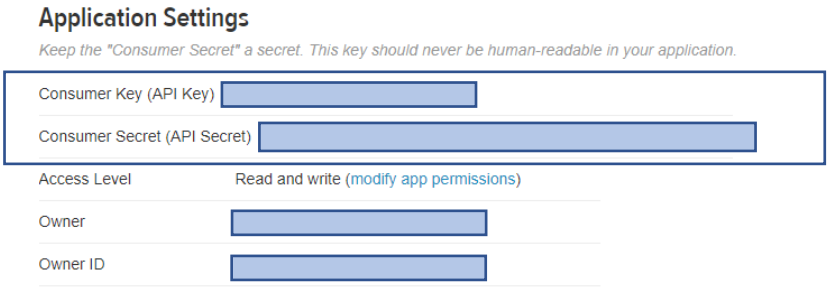
El primer paso a la hora de realizar un análisis de sentimiento es la extracción de los datos, en este caso, de los tweets. Para ello, se usará la información de Twitter que es de carácter público y a la cual se puede tener acceso de una manera sencilla.

3.1.1 Acceso a los datos de Twitter

Para la recolección de los datos textuales correspondientes a la capitalidad gastronómica de León 2018, se usará la API de Twitter. Se trata de una plataforma desde la cual esta red social proporciona la posibilidad de disponer de datos a tiempo real que pueden ser muy útiles para las empresas e instituciones.

Al crear una aplicación en dicha API, ésta proporciona los permisos necesarios para su autentificación, en la Figura 3.1 se ve que asigna dos claves; Consumer Key (API Key) y Consumer Secret (API Secret), marcadas por el rectángulo azul.

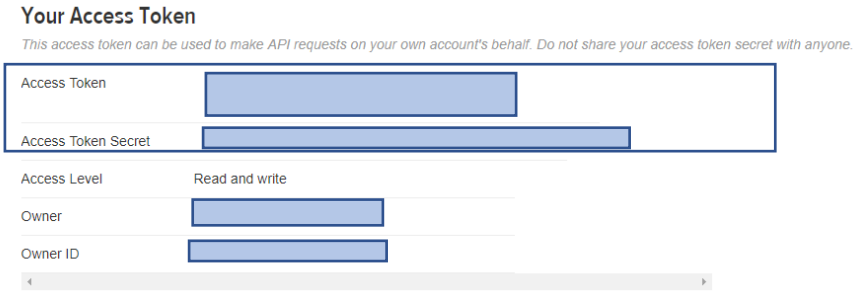
Figura 3.1 Claves para la autentización de la API



Fuente: (Twitter Application Management, s. f.)

Además, se necesitan las claves del token para poder acceder a la aplicación, la API lo muestra como se ve en la Figura 3.2. Tenemos dos claves; Access Token y Access Token Secret, que de igual manera están destacadas con el rectángulo azul.

Figura 3.2 Claves Token



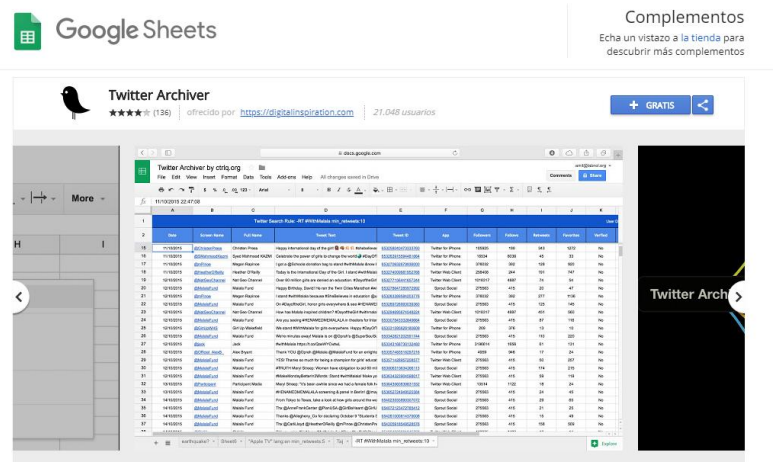
Fuente: (Twitter Application Management, s. f.)

Una vez creada la aplicación existen dos formas de extraer los datos de Twitter. Una de ellas es utilizar los paquetes de R twitteR y ROauth donde nos pedirán las claves anteriores. Y la otra forma es a través de Google Drive que es la metodología que se va a realizar para determinar el sentimiento a través de Twitter en lo referente a la capitalidad gastronómica de León y que se explicará con mayor detalle a continuación.

3.1.2 Recopilación de los datos a través de la nube

La recolección de tweets a través de Google Drive está basada en la nube. Para su realización, se accede a Google Drive y se abre una hoja de cálculo en blanco. Para poder descargar los tweets se necesita un complemento llamado Twitter Archiver que se encuentra disponible para descargar en la página de documentos de Google Sheets como muestra la Figura 3.3.

Figura 3.3. Twitter Archiver

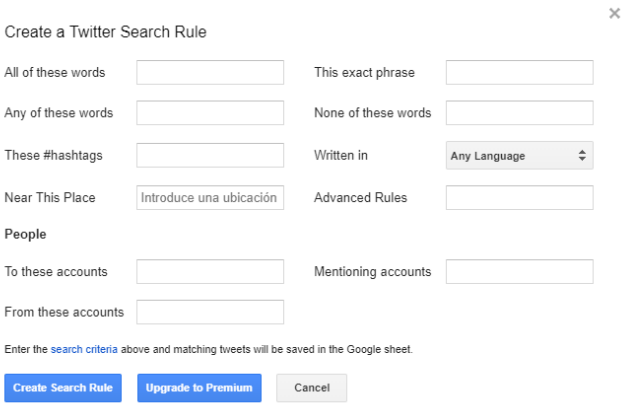


Fuente: (Google Sheets, s. f.)

Una vez instalado dicho complemento se procede a la autorización de los permisos con la cuenta de Google y ya se puede comenzar a recopilar. Twitter Archiver almacena automáticamente los tweets cada hora. Crea una base de datos con varias columnas con información sobre el tweet y sobre el usuario como: la fecha en la que se publicó el tweet, el nombre de usuario, el nombre completo del usuario, el texto del tweet, el código de identificación del tweet, la aplicación desde la cual se ha mandado el tweet, el número de seguidores del usuario, el número de favoritos, la verificación o no de la cuenta del usuario, la fecha desde la cual el usuario usa Twitter, la localización, la foto de perfil del usuario.

Además, para la recolección de los tweets a través de esta metodología existen varios filtros en función de cómo el investigador quiera obtener los datos. En la Figura 3.4 que se muestra a continuación se ve que podemos filtrar los tweets por palabras, por etiquetas, por ubicación, por perfiles, etc.

Figura 3.4 Filtros Twitter Archiver



Fuente: Twitter Archiver

Esta aplicación tiene más ventajas de recolección de tweets que los paquetes que proporciona R. Permite pedirle características específicas de cómo se quieren buscar los datos y además descarga los tweets de manera automática. Se trata de un complemento gratuito si se establece una clave de búsqueda por aplicación, en caso de querer más de una clave de búsqueda se necesitaría la versión Premium.

Para obtener los datos procedentes de Twitter que hacen referencia a la capitalidad gastronómica de la ciudad de León en el año 2018, se han filtrado los tweets de la siguiente manera:

- Utilizando el hashtag #leonesp y filtrando con las palabras: gastronomía, gastronómico, gastronómica, capital y manjar.
- Utilizando el hashtag #Leónmanjardereyes.
- Utilizando el hashtag #LeónGastro2018.

Se ha establecido una base de datos con un total de 178 tweets recolectados durante el primer semestre del año 2018, es decir, desde 01/01/2018 hasta el 15/06/2018.

3.2 PREPARACIÓN DE LA BASE DE DATOS

La fase de la preparación de la base de datos es de gran importancia. Para la realización de un análisis de sentimiento es necesario hacer una clasificación manual antes de entrenar el modelo. Se debe enseñar al ordenador las reglas necesarias para la clasificación automática de los tweets. Así, le mostraremos qué palabras están asociadas a tweets positivos y cuáles a tweets negativos. Hacer esto a base de códigos es muy complicado, el ordenador entiende mejor a base de ejemplos. Se usará machine learning para que sea capaz de comprender a base de ejemplos la clasificación de los tweets según su polaridad. Machine learning es una disciplina científica en el ámbito de la inteligencia artificial que tiene como objetivo que los sistemas informáticos aprendan automáticamente. Aprender, bajo esta perspectiva, quiere decir que se identifiquen patrones complejos extraídos de millones de datos. La máquina funciona creando un algoritmo que revisa los datos y es capaz de predecir comportamientos futuros. (González, 2014). Gracias a machine learning, una empresa o institución pasaría de tener un comportamiento activo a ser completamente proactivos ya que el algoritmo es capaz de crear patrones de comportamiento capaces de predecir el futuro.

En base a lo anterior, el lenguaje supervisado trata de aprender a base de ejemplos que le daremos al ordenador, dichos ejemplos tienen que estar clasificados manualmente por el investigador porque éstos deben estar clasificados correctamente. En general, cuantos más ejemplos clasificados le mostremos al ordenador mejor será el aprendizaje.

Se ha preparado la base de datos con la clasificación manual del sentimiento de los tweets. La base de datos cuenta con cuatro columnas, la correspondiente al texto, la correspondiente a la ubicación desde la cual se ha enviado el tweet, la correspondiente a los emojis incorporados en el tweet (en caso de que los haya) y la última que corresponde al sentimiento siendo este 1 si es positivo y -1 si es negativo.

3.2.1 Detección de la polaridad

La determinación de la polaridad tiene como objetivo detectar si el texto en cuestión es positivo o negativo. Los mensajes que proceden de las redes sociales son uno de los elementos más difíciles de clasificar por las siguientes razones:

- Los mensajes son cortos.
- Contenido ruidoso, en muchas ocasiones, los textos están mal formados en base al vocabulario, la ortografía y la sintaxis. Además, se usa un lenguaje coloquial, abreviaturas, emoticonos, alargamiento de las palabras, etc. Frente a esto, existen dos soluciones, una de ellas es adaptar y mejorar los cálculos del procesamiento de lenguaje natural para ajustarse al texto, y la otra es adaptar el texto a las tecnologías de la lengua.
- Existe un fuerte componente dinámico, característico de las redes sociales, donde hay un fuerte potencial de abrir un debate en función del tema en cuestión en un momento determinado, además, de la evolución de los temas de actualidad.
- Las redes sociales tienen un ámbito mundial por ello el investigador puede enfrentarse a mensajes escritos en diferentes idiomas.

Una vez preparada la base de datos se descarga en formato csv con separadores por comas (comma-separated-values). Se trata de un archivo de texto que almacena los datos en columnas y, las filas se distinguen por saltos de línea (Parra, 2015).

3.3 PREPROCESAMIENTO DE TEXTO

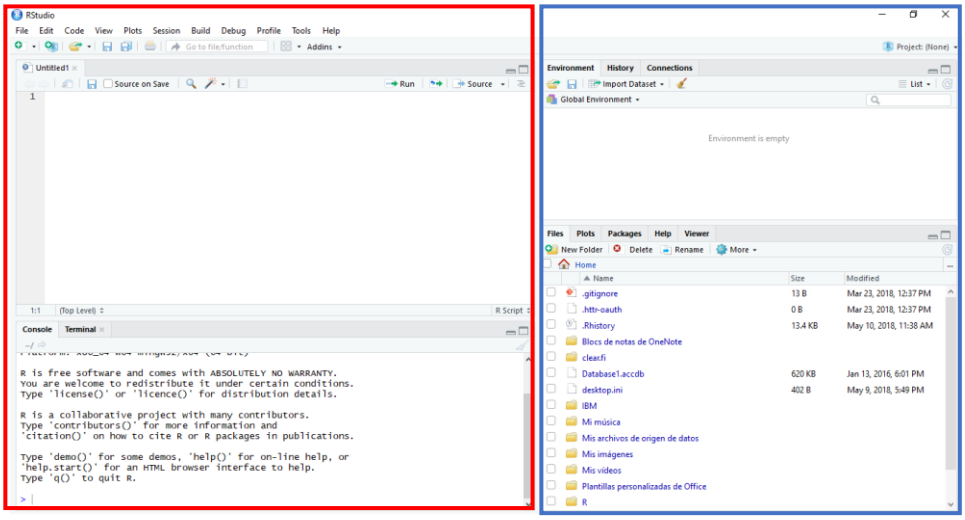
El objetivo de esta fase es preparar el texto para su posterior análisis de sentimiento, es decir, hay que simplificar los tweets para que el ordenador entienda mejor su significado. Los tweets originales tienen información que es innecesaria para el análisis como son los signos de puntuación, el uso de mayúsculas y minúsculas, espacios, etc. Todo ello se irá explicando con todo detalle a continuación.

Antes de empezar, hay que cargar la base de datos en R. R es un software de lenguaje de programación, además, de una herramienta de trabajo orientado al cálculo estadístico, manipulación de los datos y representación gráfica (JUrcera, 2012). No obstante, este trabajo se desarrollará con Rstudio que es un conjunto de herramientas integradas diseñadas para ayudar a R a ser más productivo. Incluye la consola, el resultado de la

sintaxis, un editor que permite la ejecución de código directo, así como una gran variedad de herramientas para ver el historial, depurar y administrar el espacio de trabajo.

Como se ve en la Figura 3.5, Rstudio se divide en la pantalla en cuatro partes. La parte de la derecha es donde se escribe el código de R, mostrada con un rectángulo rojo. La parte de la izquierda es la correspondiente el soporte de R, mostrada con un rectángulo azul.

Figura 3.5 Apariencia Rstudio



Fuente: Rstudio

Ahora, se carga la base de datos, previamente preparada, en Rstudio.

La siguiente instrucción indica el directorio que es la carpeta que usa R para leer ficheros:

```
getwd()
```

A continuación, se creará un objeto llamado TweetsLeon que es el que contendrá la base de datos:

```
TweetsLeon <- read.csv (“TweetsLeon.csv”, sep = ” ; ”, encoding = “ UTF-8 ”)
```

A través de la instrucción read.csv se indica a Rstudio que el documento que va a leer está en formato csv. El primer argumento es el nombre del archivo que se denomina TweetsLeon, el segundo argumento es el separador por comas, y el tercer argumento es el código de codificación de caracteres UTF-8 para que no haya problemas con los posibles diferentes idiomas de los tweets.

Antes de comenzar a trabajar es conveniente introducir la siguiente línea de código:

```
TweetsLeon$Texto <- chartr (“áéíóúñ”, “aeiouñ”, TweetsLeon$Texto)
```

Al tener los tweets en español es posible que haya problemas con el uso de las tildes y las ñ, es por ello, que es muy aconsejable eliminarlas del texto para que el análisis sea más efectivo.

Una vez cargada la base de datos, es interesante saber cuántos tweets son positivos y cuántos son negativos. Para ello, usaremos la siguiente función:

```
table (TweetsLeon$Sentimiento)
```

El resultado se muestra en la Tabla 3.1:

Tabla 3.1 Recuento sentimiento tweets León

RECUENTO TWEETS LEÓN	
-1	1
20	158

Fuente: Elaboración Propia

No obstante, se ha creado un gráfico para la mejor visualización de los datos. Se ha usado el paquete de R ggplot2:

```
Install.packages (“ggplot2”)

Library (“ggplot2”)
```

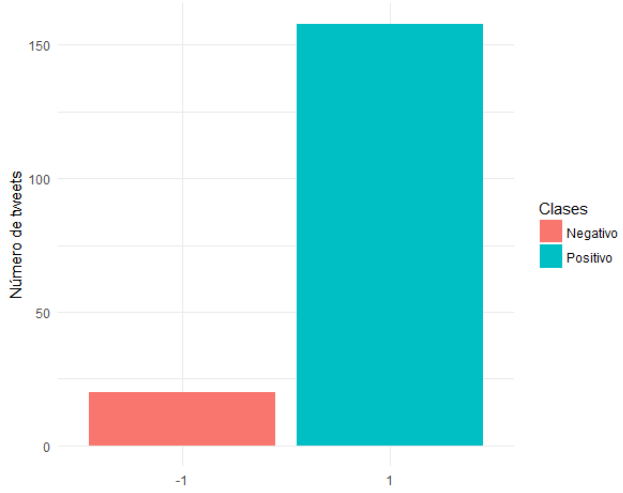
El paquete ggplot2 es un sistema para crear gráficos, basado en la “gramática de los gráficos” que se refiere a especificar de forma independiente cada componente del grafico para después combinarlos.

Para la realización del gráfico de barras se ha escrito la siguiente línea de código (Rstudio, s. f.):

```
Qplot (data=TweetsLeon, factor (TweetsLeon$sentimiento), geom= "bar", ylab= "Número de tweets", xlab= "", fill= factor (TweetsLeon$sentimiento)) + theme_minimal () + scale_fill_discrete (name= "Clases", labels= c("Negativo", "Positivo"))
```

El resultado se muestra en el Gráfico 3.2:

Gráfico 3.2 Sentimiento Tweets León



Fuente: Elaboración Propia a partir del paquete ggplot2

Para comenzar con el preprocesamiento de texto hay que instalar y cargar en Rstudio dos paquetes:

```
Install.packages ("tm")

Library ("tm")

Install.packages ("SnowballC")

Library ("SnowballC")
```

El paquete tm quiere decir text mining que corresponde al paquete para la minería de texto de R. Ofrece funcionalidad para gestionar documentos de texto, manipulación de dichos documentos y facilita el uso de formatos heterogéneos de texto (tm.r-forge.r-project.org, s. f.).

El paquete SnowballC lleva implementado un algoritmo que permite realizar el steming, es decir, reducir las palabras a su raíz para ayudar a la comparación del vocabulario. Los idiomas que lleva incorporados actualmente son: danés, holandés, inglés, finlandés, alemán, húngaro, italiano, noruego, portugués, rumano, ruso, español, sueco y turco. (Bouchet-Vala, 2015).

Ahora, es el momento de empezar a trabajar con los datos.

3.3.1 Creación del Corpus

En el estudio de una lengua es esencial trabajar con un corpus cuyo objetivo es analizar y conocer el uso de determinadas palabras que componen una oración, así como saber la frecuencia del uso de dichas palabras. En definitiva, un corpus es un almacén de datos (numéricos o textuales) de las mismas características que sirven como base de una investigación, en este caso el corpus estará formado por los tweets previamente recolectados. Gracias a este corpus Rstudio va a poder trabajar con los tweets porque los va a entender.

Para crear el corpus en Rstudio se escribe la siguiente línea de código:

```
CorpusLeon <- Corpus (VectorSource (TweetsLeon$Texto), readerControl = list
(readPlain, lenguajeE1 = "es", load = TRUE))
```

Con el código anterior R leerá los tweets dentro de un corpus, lo que permitirá trabajar con los datos y manipular cada una de las palabras que contienen dichos tweets. A través de la función VectorSource se convierte cada tweet en un documento en forma de lista, además se le indica que el idioma en cuestión es el español.

Para comprobar que el corpus se ha formado de una manera correcta, usaremos la siguiente función para ver cuántas filas hay en el corpus:

```
Lenght (CorpusLeon)

[1] 178
```

El resultado nos muestra un total de 178 tweets que son los que efectivamente se han incorporado, por lo tanto, el corpus se ha creado de forma correcta.

El siguiente paso es comprobar si los tweets están de manera correcta dentro del corpus, para ello se usa la siguiente función:

```
Content (CorpusLeon [50])

[1] "Se nos hace la boca agua con estos lomos de sardina ahumadas con pimientos.
#leonesp #gastronomia"
```

Lo que se le pide es que muestre el documento 50, y comprobamos que el texto del tweet está de manera correcta sin errores.

Una vez que se ha comprobado que el corpus funciona correctamente, es el momento de comenzar el preprocesamiento de texto con sus diferentes fases.

3.3.2 Limpieza general

En esta primera fase del preprocesamiento de texto el objetivo es convertir todas las palabras a minúsculas, eliminar los signos de puntuación, los números que pueda haber en el texto y los posibles espacios de sobra.

Para ello, se modifica el objeto CorpusLeon con la función tm_map que sirve para realizar transformaciones. Tiene como primer argumento el nombre del corpus y como segundo argumento el preprocesamiento que se desea hacer, en este caso, es convertir todas las palabras a minúscula:

```
CorpusLeon <- tm_map (CorpusLeon, tolower)
```

Al correr este código, el tweet anterior que sirvió para comprobar el corpus debería haber cambiado, para ello, usamos la función content y se ve que efectivamente está en minúsculas.

```
Content (CorpusLeon [50])
```

```
[1] "se nos hace la boca agua con estos lomos de sardina ahumadas con pimientos.
#leonesp #gastronomia"
```

Una vez que se ha convertido el corpus a letras minúsculas, se procede a eliminar los signos de puntuación con la siguiente función:

```
CorpusLeon <- tm_map (CorpusLeon, removePuntuaction)
```

Del mismo modo, usamos la función content para ver el tweet 50 y se comprueba que efectivamente se ha eliminado la puntuación.

```
Content (CorpusLeon [50])
```

```
[1] "se nos hace la boca agua con estos lomos de sardina ahumadas con pimientos
leonesp gastronomia"
```

La siguiente función eliminará los números del texto de los tweets, en caso de que los tengan:

```
CorpusLeon <- tm_map (CorpusLeon, removeNumbers)
```

También, es necesario eliminar los posibles espacios en blanco sobrantes:

```
CorpusLeon <- tm_map (CorpusLeon, stripWhitespace)
```

Una vez que se ha realizado la limpieza general del corpus, se crea una nube de palabras para observar los términos más frecuentes del CorpusLeon.

Se entiende por nube de palabras una representación gráfica de las palabras más usadas en un texto, siendo las palabras más frecuentes las de tamaño mayor en la representación.

Para la realización de la nube de palabras en Rstudio hay que instalar y cargar dos paquetes

```
install.packages ("wordcloud")

library ("wordcloud")

install.packages ("RColorBrewer")

library ("RColorBrewer")
```

El paquete wordcloud tiene como objetivo generar un gráfico en forma de nube comparando las frecuencias de las palabras de los documentos textuales de una base de datos (Fellows, 2015).

El paquete RColorBrewer proporciona esquemas de color para mapas y gráficos en R (Neuwirth, 2015).

```
Wordcloud (CorpusLeon, random.order = FALSE, colors = brewer.pal (8, "Dark2"), max.words = 100)
```

Con la anterior línea de código se crea una nube de palabras del objeto CorpusLeon, la función random.order = FALSE indica que el orden en el que aparezcan las palabras no va a ser aleatorio, las palabras más mencionadas aparecerán en el centro del gráfico. El código colors dará una combinación de colores para que las palabras cambien su color en función de la frecuencia de éstas. Además, se ha establecido un máximo de 100 palabras y cuyo resultado se muestra en el Cuadro 3.1. Se observa que las palabras más repetidas en el corpus son "león", "leonesp", "gastronomía", "capital", "manjar", etc. No es de

extrañar porque se trata de las palabras usadas para la recolección. No obstante, también se ven palabras como "que", "por", "con", "los", etc. En base a esto es necesario continuar con la limpieza para eliminar todo este "ruido" que no aporta significado alguno.

Cuadro 3.1 Nube de palabras CorpusLeon

Fuente: Elaboración Propia a través del paquete Wordcloud

3.3.3 Stop word removal

El objetivo de esta fase es eliminar las palabras más frecuentes para la lengua. Son los pronombres, artículos y demás palabras que no aportan mucho valor al sentimiento del tweet, por ello, es conveniente eliminarlos. R tiene precargadas las stop words en cada idioma. Para verlas se escribe la siguiente línea de código, en este caso el idioma es el español:

```
stopwords ("spanish")
```

El resultado muestra 309 palabras que se eliminarán en el momento que R las detecte en el texto del tweet.

Para eliminar dichas palabras del corpus se escribe la siguiente función:

```
CorpusLeon <- tm_map (CorpusLeon, removeWords, c(stopwords ("spanish")))
```

Ahora, se comprueba que las ha eliminado:

```
Content (CorpusLeon [50]

[1] "hace boca agua lomos sardina ahumadas pimientos leonesp gastronomia"
```

3.3.4 Steming

Es la última fase del preprocesamiento de texto. El objetivo es cortar las palabas a su raíz, por ejemplo, la palabra “boca” se convertiría en “boc” y todas las palabras que tengan como raíz “boc” serían la misma. Es una manera de simplificar los datos para optimizar el análisis de sentimiento.

Para su realización en Rstudio, se escribe la siguiente función:

```
CorpusLeon <- tm_map (CorpusLeon, stemDocument)
```

Comprobamos que ha reducido las palabras a su raíz:

```
Content (CorpusLeon [50])

[1] “hac boc agu lom sardin ahum pimiento leonesp gastronomi”
```

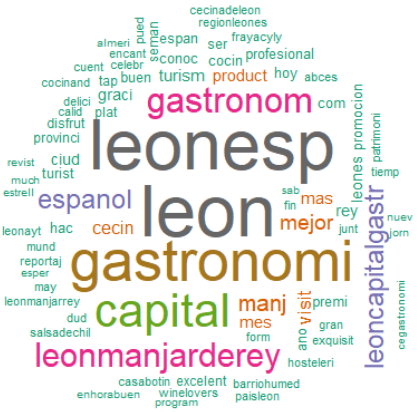
Gracias a esto, muchas palabras se tratarán de manera homogénea y se evitaran problemas en cuanto al procesamiento de la información (Martínez Gordillo, 2016).

Ahora se genera una nube de palabras para ver de forma gráfica aquellas que tienen mayor frecuencia en el corpus una vez reducidas a su raíz:

```
Wordcloud (CorpusLeon, random.order = FALSE, colors = brewer.pal (8, "Dark2"), max.words = 80)
```

A través de la función wordcloud se genera la nube de palabras del objeto CorpusLeon una vez hecho el steming y se le pide un máximo de palabras igual a 80. El resultado se muestra en el Cuadro 3.2. Se ve que ha habido cambios en cuanto al Cuadro 3.1, ya que las palabras están reducidas a su raíz y se han eliminado las stopwords por lo que se ha simplificado el corpus de manera sustancial. No obstante, siguen apareciendo palabras que no aportan significado.

Cuadro 3.2 Nube de palabras CorpusLeón stem



Fuente: Elaboración Propia a través del paquete Wordcloud

En base al resultado anterior, se pueden eliminar más palabras del corpus ya que como se observa, muchas palabras que aparecen reflejadas pueden aparecer tanto en tweets positivos como en tweets negativos, además, también, como ya se ha dicho, son palabras que han sido usadas para la recolección de los tweets por lo que su frecuencia es muy alta. Por lo tanto, se procederá a eliminar las siguientes palabras del corpus: “leonesp”, “león”, “leonmanjarderey”, “gastronom”, “gastronomi”, “manj”, “capital”, “espanol”, “leoncapitalgastr”.

En definitiva, no son palabras determinantes para la clasificación de los tweets.

```
CorpusLeon <- tm_map (CorpusLeon, removeWords, c("leonesp", "leon", "leonmanjarderey", "gastronom", "gastronomi", "manj", "capital", "espanol", "leoncapitalgastr"))
```

Para ver los cambios gráficamente se realiza una nube de palabras una vez eliminadas las palabras anteriores:

```
Wordcloud (CorpusLeon, random.order = FALSE, colors = brewer.pal (8, "Dark2"), max.words = 70)
```

El resultado se muestra en el Cuadro 3.3. Se observa que ahora es posible hacerse una idea de las palabras más usadas en el corpus. Se ve que la palabra “mejor” es la más grande por tanto su frecuencia es alta, también palabras como “cecina” y “visitar”.

Después también son relevantes palabras como “profesional”, “excelente”, “premio”, “producto”, “promoción”, “disfrutar”, “turismo”, etc.

Cuadro 3.3 Nube de palabras CorpusLeon stem 2

Fuente: Elaboración Propia a través del paquete Wordcloud

3.3 ANÁLISIS DE SENTIMIENTO

3.3.1 Extracción de características

En este apartado se va a explicar la manera en la que los tweets se van a almacenar en la base de datos. Para ello, se va a poner un ejemplo; el tweet que se ha escogido al azar anteriormente para comprobar el preprocesamiento de texto y una vez finalizado éste era: “hac boc agu lom sardin ahum pimient”.

El ordenador es capaz de realizar modelos estadísticos a partir de números o de variables binarias, pero es muy complicado realizar dichos modelos únicamente a partir de palabras (Martínez Gordillo, 2016). Por tanto, hay que expresar y convertir ese texto en stem a números. Para la conversión de las palabras a números se genera una base de datos donde cada uno de los tweets componen las filas y las columnas serán las variables que son todas las palabras posibles que puedan aparecer en dichos tweets, es decir, habrá tantas columnas como palabras haya en todos los tweets.

Por lo tanto, ante el tweet “hac boc agu lom sardin ahum pimient”, se generaría la siguiente base de datos representada en la Tabla 3.2:

Tabla 3.2 Extracción de características

Sentimiento	hac	boc	agu	sardin	ahum	pimient
1	1	1	1	1	1	1

Fuente: Elaboración Propia

Para rellenar la base de datos en su totalidad, se pondría un 1 si la palabra en cuestión aparece en el tweet y un 0 en caso de que no apareciera. De este modo, la base de datos total tendría un aspecto parecido a la base de datos que se representa en la Tabla 3.3:

Tabla 3.3 Base de datos total

Tweet	Sentimiento	Palabra 1	Palabra 2	…	Palabra n
1	-1	1	0	0	…
2	-1	0	1	0	…
3	1	0	1	1	…
…	…	…	…	…	…
n	…	…	…	…	…

Fuente: Elaboración Propia basado en (Martínez Gordillo, 2016)

3.3.1.1 Creación de la matriz

Una vez finalizado el preprocesamiento de texto, es el momento de crear la matriz que se ha explicado anteriormente. Para ello, se ha asignado la siguiente función al objeto denominado frequenciesLeon:

```
FrequenciesLeon <- DocumentTermMatrix (CorpusLeon)
```

Una vez creada la matriz de texto, se inspecciona:

```
FrequenciesLeon

<<DocumentTermMatrix (documents: 178, terms: 909)>>

Non-/sparse entries: 1396 / 160106

Sparsity: 99%

Maximal term lenght: 46
```

```
Weighting: term frequency (tf)
```

Esta información que muestra R dice que FrequeciesLeon es una matriz de documentos que tiene 178 documentos, con 909 términos, es decir, con 909 columnas. El largo máximo de todas las columnas es la palabra más larga que hay en la matriz y que es de 46 caracteres. En definitiva, la matriz creada llamada FrequenciesLeon tiene un tamaño de 178 filas x 909 columnas.

La matriz que se ha creado, denominada FrequenciesLeon, indica también una dispersión del 99% lo que indica que está casi vacía debido a la cantidad de 0 que tiene. Esto es debido a que hay palabras que solo se mencionan una vez en un tweet determinado.

Ya se ha creado la matriz de forma correcta, es interesante ver ahora cómo se ha rellenado la matriz FrequenciesLeon. Para ello, se va a hacer un "zoom" de dicha matriz:

```
Inspect (FrequenciesLeon [50:55, 60:65])
```

Con esta función se le pide a R que muestre de la fila 50 a la 55 y de la columna 60 a la 65. El resultado que se muestra en la Tabla 3.4 que está rellenada en su mayoría con 0 porque en los tweets seleccionados no aparecen las palabras correspondientes a las columnas 60-65, excepto en los tweets número 54 y 55 que aparece la palabra promoción y por ello, las casillas correspondientes tienen un 1

Tabla 3.4 Matriz FrequenciesLeon [50:55, 60:65]

Terms / Docs	asi	camarer	promocion	public	subvencion	supuest
50	0	0	0	0	0	0
51	0	0	0	0	0	0
52	0	0	0	0	0	0
53	0	0	0	0	0	0
54	0	0	1	0	0	0
55	0	0	1	0	0	0

Fuente: Elaboración Propia

3.3.1.2 Frecuencia de las palabras

El siguiente paso es determinar cuáles son los términos más frecuentes de la base de datos en cantidades. Para ello, se transforma el objeto FrequenciesLeon en una matriz denominada FrequenciesLeonMatrix que tendrá como filas los tweets (178), y como columnas el total de palabras que hay en el corpus (909):

```
FrequenciesLeonMatrix <- as.matrix (FrequenciesLeon)
```

Para obtener la frecuencia de las palabras se suman las columnas que, como ya se ha dicho anteriormente, están compuestas por 0 en caso de que la palabra no esté en el tweet o 1 en caso de que sí esté. Por lo tanto, al realizar dicha suma se obtendrá las frecuencias de las palabras para saber cuáles son las más usadas y en qué medida.

```
Library ("magrittr")

FrequenciesLeonMatrix <- colsums (FrequenciesLeonMatrix) %>% sort (decreasing = TRUE)
```

En la anterior línea de código, aparece el operador %>% que pertenece a la librería magrittr. Funciona como una "tubería" y se usa para "encadenar" funciones, lo que conlleva el mismo funcionamiento que la anidación (Milton Bache y Wickham, 2014). Lo que se pide es que se sumen las columnas de la matriz FrequenciesLeonMatrix y que se ordenen de mayor a menor para conocer la frecuencia de cada palabra ordenando de mayor a menor.

Posteriormente, se guardan los resultados en un data frame. Un data frame es una colección de variables que se utiliza como la estructura de datos fundamental por la mayoría de software de modelado de R (R: Data Frames, s. f.). Dicho data frame está compuesto por dos columnas, una correspondiente al nombre de la palabra y otra que corresponde a la frecuencia:

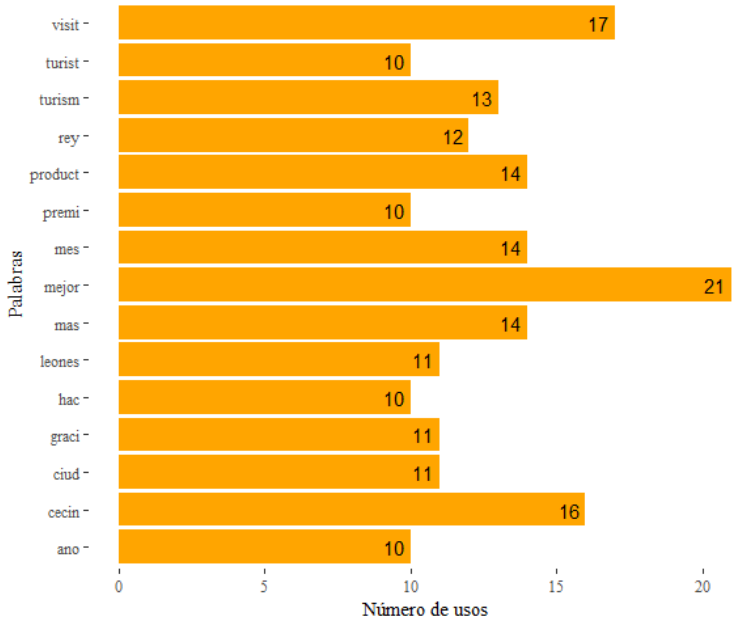
```
FrequenciesLeonMatrix <- data.frame (palabra = names (FrequenciesLeonMatrix), freq = FrequenciesLeonMatrix)
```

Para hacer más visuales los resultados, se crea un gráfico utilizando la librería ggplo2, utilizando un tema adicional incluido en el paquete ggtheme que incorpora temas y escalas adicionales al paquete ggplot2 :

```
Install.packages ("ggtheme")

Library ("ggtheme")

FrequenciesLeonMatrix [1:15, ] %>% ggplot (aes (palabra, freq)) + geom_bar (stat = "identity", fill = "Orange") + geom_text (aes (hjust = 1.3, label = freq)) + coord_flip() + labs ( x = "Palabras", y = "Número de usos") + theme_tufte()
```

Para generar el gráfico, se ha usado la línea de código anterior que a través de la matriz FrequenciesLeonMatrix, usando las 15 palabras más usadas, se genera el Gráfico 3.3. Se ve que la palabra más frecuente, como ya se mencionaba en la nube de palabras, es "mejor" con 21 repeticiones, le sucede "visit" con 17 repeticiones y "cecin" con 16. Además, están las palabras "product", "premi", "graci", etc.

Gráfico 3.3 Palabras más frecuentes en Tweets León



Fuente: Elaboración Propia a partir del paquete ggplot2

3.4.1.3 Reducción de la matriz

Existen muchas palabras que no son relevantes para el análisis debido a su baja frecuencia. Lo que se necesita es establecer patrones para que a partir de ellos el ordenador pueda entender el sentimiento. Por lo tanto, todas esas palabras que no se repiten mucho se eliminarán. Para ello, se creará una nueva matriz denominada SparseLeon:

```
SparseLeon <- removeSparseTerms (FrequenciesLeon, 0.98)
```

La matriz SparseLeon está basada en la matriz FrequenciesLeon. Se establece un umbral de 0.98 lo que indica que un término tiene que aparecer en más de un 2% de los documentos para no ser eliminado.

Para obtener más información sobre esta eliminación:

```
SparseLeon

<<DocumentTermMatrix (documents: 178, terms: 82)>>

Non-/sparse entries: 548 / 140448

Sparsity: 96%

Maximal term lenght: 16

Weighting: term frequency (tf)
```

Muestra que la matriz contiene el mismo número de documentos, 178, pero el número de términos ha disminuido. Hay 82 palabras que son las más usadas, por lo tanto, esas son las palabras que interesan para el análisis. También se observa que la dispersión se ha reducido de un 99% a un 96%.

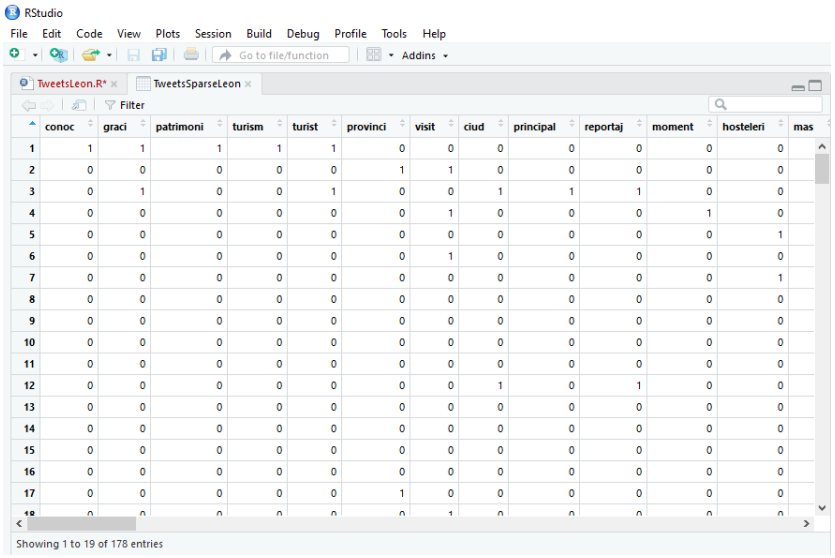
Una vez creada la matriz reducida, es necesario convertir dicha matriz de términos a un data frame en Rstudio para poder trabajar un modelo.

Se crea un data frame llamado TweetsSparseLeon basado en la matriz SparseLeon:

```
TweetsSparseLeon <- as.data.frame (as.matrix (SparseLeon))
```

Lo que va a hacer esta función es convertir la matriz SparseLeon en una base de datos en formato R. En la Tabla 3.5 se ve el contenido de dicha base de datos se comprueba que efectivamente está rellenada por 0 y 1.

Tabla 3.5 Data frame TweetsSparseLeon



	conoc	graci	patrimoni	turism	turist	provinci	visit	ciud	principal	reportaj	moment	hosteleri	mas
1	1	1	1	1	1	0	0	0	0	0	0	0	
2	0	0	0	0	0	1	1	0	0	0	0	0	
3	0	1	0	0	1	0	0	1	1	1	0	0	
4	0	0	0	0	0	0	1	0	0	0	1	0	
5	0	0	0	0	0	0	0	0	0	0	0	1	
6	0	0	0	0	0	0	1	0	0	0	0	0	
7	0	0	0	0	0	0	0	0	0	0	0	1	
8	0	0	0	0	0	0	0	0	0	0	0	0	
9	0	0	0	0	0	0	0	0	0	0	0	0	
10	0	0	0	0	0	0	0	0	0	0	0	0	
11	0	0	0	0	0	0	0	0	0	0	0	0	
12	0	0	0	0	0	0	0	1	0	1	0	0	
13	0	0	0	0	0	0	0	0	0	0	0	0	
14	0	0	0	0	0	0	0	0	0	0	0	0	
15	0	0	0	0	0	0	0	0	0	0	0	0	
16	0	0	0	0	0	0	0	0	0	0	0	0	
17	0	0	0	0	0	1	0	0	0	0	0	0	

Showing 1 to 19 of 178 entries

Fuente: Elaboración Propia a partir de Rstudio

Una vez creada esta base de datos falta algo muy importante que es la variable correspondiente al sentimiento. En esta nueva matriz y base de datos no existe dicha variable, por lo tanto, hay que incorporársela de alguna manera. Para ello, se escribe la siguiente línea de código que va a asignar la variable sentimiento de la base de datos inicial TweetsLeon al data frame que se ha creado TweetsSparseLeon:

```
TweetsSparseLeon$sentimiento <- TweetsLeon$sentimiento
```

3.3.2 Creación del modelo de clasificación

Una vez que ha sido creado el data frame con la matriz final ya se puede realizar el modelo de clasificación. El modelo de clasificación que se va a usar es suport vector machine de machine learning. La clasificación de los datos de forma automática pertenece a la ciencia de la computación relacionada con la capacidad de las máquinas de aprender, además de ser capaces de reconocer patrones (pattern recognition), es decir, la capacidad del

ordenador de aprender sin ser programados explícitamente para ello (Madroñal Quitín, 2015).

3.3.2.1 Support Vector Machine

El método de clasificación-regresión, llamado máquinas de vector de soporte (support vector machine SVM), fue desarrollado en los años 90 dentro del campo de la ciencia computacional. Support vector machine se considera uno de los mejores clasificadores dentro de un amplio abanico de aplicaciones. Es por ello por lo que está considerado como uno de los principales referentes dentro del ámbito del aprendizaje estadístico y machine learning.

Una máquina de vectores de soporte es una herramienta de clasificación discriminatoria que está definida por un hiperplano de separación. El objetivo es generar un modelo de clasificación que, después del aprendizaje, sea capaz de distinguir automáticamente las clases a las que pertenecen los datos.

Este modelo de clasificación se realiza mediante una etapa denominada entrenamiento supervisado (supervised learning). En esta etapa se introduce una base de datos que ha sido previamente analizada y clasificada. Así, el clasificador aprenderá y, por tanto, será capaz de realizar una distinción entre las clases usando los patrones que han sido extraídos en dicha la fase de entrenamiento.

En la segunda etapa, el clasificador debe ser capaz de distinguir de un manera correcta y automática los nuevos datos de entrada que no pueden haber sido usados en la fase de entrenamiento.

3.3.2.2 Hiperplano de separación, maximal margin clasiffer

Las máquinas de vectores de soporte tienen su base fundamental en el concepto de hiperplano de separación que define los límites de decisión. Se entiende por hiperplano de separación aquel que separa un conjunto de objetos con características de diferentes clases. En un espacio p-dimensional, un hiperplano es definido como un subespacio plano y afín, es decir, que no tiene la obligación de pasar por el origen, de dimensiones p-1. En

un espacio de dos dimensiones, el hiperplano consta de un subespacio de una dimensión, por lo tanto, se trata de una recta.

La definición matemática de este hiperplano de separación para dos dimensiones se describe mediante la ecuación de la recta:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0$$

Los parámetros β_0, β_1, β_2, todos los pares de valores $x = (x_1, x_2)$ para los que se cumple la igualdad son puntos del hiperplano.

Para el caso de p-dimensiones, la ecuación se generaliza:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p = 0$$

Del mismo modo, todos los puntos definidos por el vector $x = (x_1, x_2, \ldots, x_n)$ que cumplen la ecuación pertenecen al hiperplano.

3.3.2.3 Clasificación binaria con un hiperplano de separación

Se dispone de n observaciones, cada una con p predictores y cuya variable dependiente tiene dos niveles (-1 negativo, 1 positivo). En base a esto, es posible construir un hiperplano que permita clasificar y predecir a qué grupo pertenece una observación en función de sus predictores.

Si la distribución de las observaciones permite que se puedan separar de manera lineal en dos clases (-1, 1), un hiperplano de separación cumple:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p > 0, si\ \gamma_1 = 1$$

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p < 0, si\ \gamma_1 = -1$$

Al definir cada clase como 1 o -1, y sabiendo que multiplicar dos valores negativos resultan un valor positivo, las condiciones anteriores se pueden simplificar de la siguiente manera:

$$\gamma_i(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p) > 0, para\ i = 1 \ldots n$$

En este contexto, este clasificador tiene el objetivo de asignar cada observación a una clase dependiendo del lado del hiperplano en el que se encuentre. Por ejemplo, la observación x^* se clasifica acorde a la función $f(x^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_p x_p^*$. Si $f\ (x^*)$ es positiva, la observación será asignada a la clase 1, si es negativa será asignada a la clase -1. Además, la magnitud de $f\ (x^*)$ permite determinar la distancia entre la observación y el hiperplano para conocer la confianza de la clasificación.

Para el caso de observaciones separables linealmente en dos clases, resulta un numero infinito de posibles hiperplanos, por lo que es necesario seleccionar uno de ellos como clasificador óptimo.

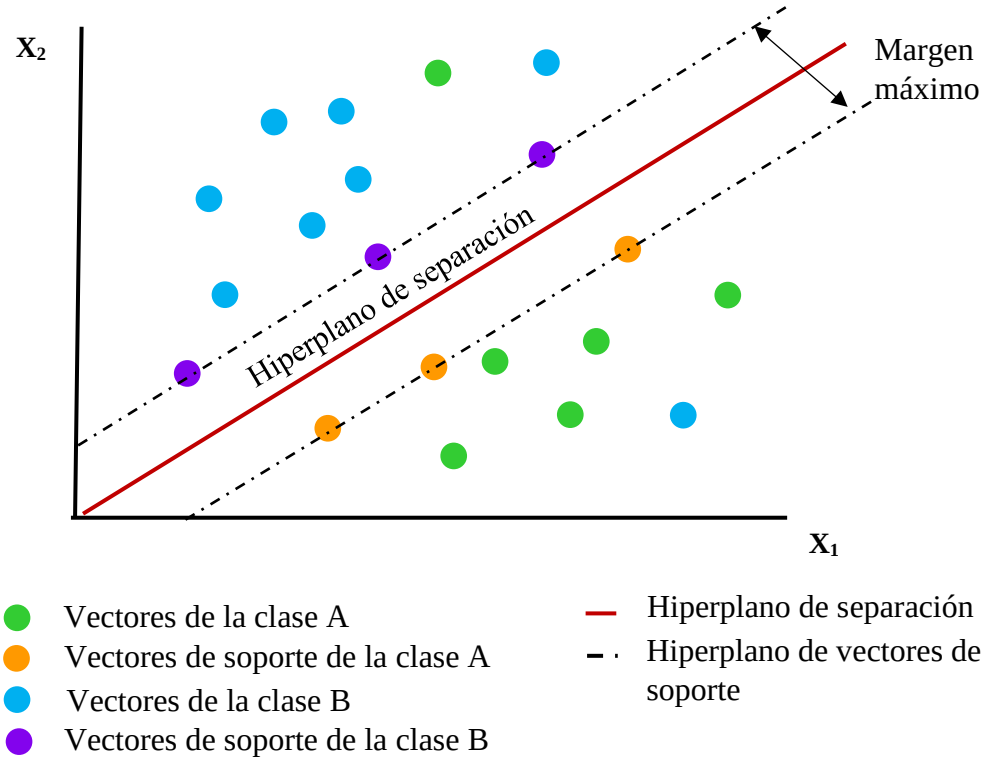
3.3.2.4 Margen máximo a partir de los vectores de soporte

En definitiva, la base de datos se encuentra en un espacio n-dimensional (donde n es igual al número de características o features que se hayan extraído) en tantos subespacios como clases, además, se otorga a cada subespacio el máximo margen de seguridad posible, es decir, la distancia que hay entre el hiperplano de separación y el punto más cercano al mismo.

En support vector machine, para determinar la línea de separación o hiperplano no busca calcular la función de densidad de probabilidad de cada clase, sino que el hiperplano de separación está determinado a partir de las muestras de entrenamiento. La característica específica de support vector machine es que para calcular el hiperplano de separación de las clases se tienen en cuenta solamente un número limitado de muestras del conjunto de entrenamiento con unas propiedades determinadas. A estas muestras de gran importancia se les denomina vectores de soporte, así cada clase tendrá un conjunto de vectores de soporte que son elegidos de manera que la distancia entre los planos, es decir, el margen sea máxima. Esta condición de margen máximo implica que se encuentra la región más amplia del espacio de características que separa las clases y que está vacía de muestras. Esta región está definida por los dos hiperplanos de vectores de soporte que son los que contienen los vectores de soporte de cada clase, por tanto, el plano intermedio de esta región es el hiperplano de separación de support vector machine.

En la Figura 3.6, los objetos pertenecen a dos clases diferentes, la línea de separación establece que los objetos que están a la derecha son verdes y pertenecen a la clase A, y los que están a la izquierda son azules y pertenecen a la clase B. En base a esto, cualquier objeto que se sitúe a la derecha de la línea se clasificará como verde y si se sitúa a la izquierda se clasificara como azul. También se puede observar que hay puntos mal clasificados porque esta herramienta de clasificación tiene un margen de error.

Figura 3.6 Representación Support Vector Machine



Fuente: Elaboración Propia basado en (Amat Rodrigo, 2017)

Cada punto en el espacio corresponde a un tweet y la ubicación de ese punto está determinada por las palabras que componen dicho tweet. Lo que busca este algoritmo de machine learning es encontrar un hiperplano capaz de dividir las clases.

3.3.2.5 Conjunto de entrenamiento y de evaluación

Para ver cómo se comporta el algoritmo de clasificación en la base de datos TweetsSparseLeon lo primero que hay que hacer es definir un conjunto de entrenamiento y otro de evaluación.

Para definir estos dos conjuntos, tal y como se muestra en la Figura 3.7, hay que dividir la base de datos: un 80% para el entrenamiento y un 20% para la evaluación. Con la parte correspondiente al entrenamiento lo que se va a hacer es entrenar el modelo, es decir, enseñarle al algoritmo de support vector machine cómo debe clasificar. Con la parte correspondiente a la evaluación se va a examinar el modelo, es decir, comprobar el poder predictivo de éste una vez que ha sido generado.

Figura 3.7 Partición de la base de datos

Fuente: Elaboración Propia

Para la división de la base de datos hay que instalar y cargar en Rstudio el paquete caTools:

```
Install.packages ("caTools")

Library ("caTools")
```

Se trata de un paquete que contiene varias funciones básicas de gran utilidad como funciones estadísticas de ventana en movimiento, lectura y/o escritura de archivos binarios, etc. (Tuszynski, s. f.). Será de gran ayuda para hacer esta partición de los datos.

Para comenzar, hay que establecer un punto de aleatoriedad para dividir la base de datos en sus correspondientes conjuntos de entrenamiento y evaluación de una manera aleatoria. Para ello, hay que establecer una "semilla" en Rstudio, de la siguiente manera:

```
Set.seed (12)
```

La base principal para usar una semilla es poder reproducir una secuencia particular de números aleatorios. La semilla propiamente dicha no tiene ningún significado inherente, sino que es una manera de decirle al generador de números aleatorios por dónde empezar. El número de la semilla que se elige es el punto de partida utilizado en la generación de

una secuencia de números aleatorios, es por ello, que utilizando siempre el mismo generador de números aleatorios y estableciendo el mismo número de semilla, siempre saldrán los mismos resultados (Cross Validated, 2018). En este caso, se ha elegido el número 12 de forma aleatoria.

Una vez establecido el punto de aleatoriedad, se creará una nueva variable llamada SplitLeon que va a definir las observaciones que corresponden al conjunto de entrenamiento y las observaciones que corresponden al conjunto de evaluación:

```
SplitLeon <- sample.split (TweetsSparseLeon$sentimiento, SplitRatio = 0.8)
```

Esta función sample.split tiene como objetivo la partición de la base de datos en subconjuntos. Toma como primer argumento la variable sentimiento correspondiente a la base de datos TweetsSparseLeon que es donde se quiere hacer la división de una manera equitativa y aleatoria, como segundo argumento tiene el ratio de separación igual a 0.8, es decir el 80% de los tweets que corresponden al conjunto de entrenamiento.

Al correr el código, la variable muestra las observaciones de la siguiente manera:

```
SplitLeon

[1]  TRUE TRUE TRUE FALSE TRUE TRUE TRUE FALSE FALSE …
```

Ahora, hay que rearmar las bases:

```
TrainSparseLeon <- subset (TweetsSparseLeon, SplitLeon == TRUE)
```

Con esta función se crea un subconjunto de datos dentro de la base de datos TweetsSparseLeon compuesto por las observaciones del SplitLeon igual a TRUE, así se ha creado el conjunto de entrenamiento.

A continuación, del mismo modo, se creará el conjunto de evaluación:

```
TestSparseLeon <- subset (TweetsSparseLeon, SplitLeon == FALSE)
```

Así se crea el otro subconjunto de la base de datos TweetsSparseLeon compuesto por las observaciones del SpliLeon igual a FALSE, que corresponde al conjunto de evaluación.

Se comprueba que se ha generado la división de forma correcta asignando un 80% de los tweets al conjunto de entrenamiento y un 20% de los tweets al conjunto de evaluación:

- TrainSparseLeon tiene un total de 142 observaciones.
- TestSparseLeon tiene un total de 36 observaciones.

Ya se han creado los dos subconjuntos (entrenamiento y evaluación), es el momento de crear el algoritmo de support vector machine. Para ello, hay que instalar y cargar tres paquetes de Rstudio:

```
Install.packages ("caret")

Library ("caret")

Install.packages ("e1071")

Library ("e1071")

Install.packages ("plyr")

Libray ("plyr")
```

El paquete de R caret (classification and regression training) contiene una serie de funciones que facilitan el uso de muchos métodos complejos de clasificación y regresión (Berrendero, 2017). En definitiva, se trata de un paquete que contiene funciones diversas para el entrenamiento y la clasificación de modelos de regresión (Max Kuhn Contributions from Jed Wing et al., 2018).

El paquete de R, e1071, contiene las funciones necesarias, entre otras, para crear el algoritmo support vector machine de machine learning (Mayer et al., 2017).

El paquete plyr tiene como principales funciones dividir y combinar los datos (Wickham y Maintainer, 2016).

Con estos tres paquetes se puede crear el modelo de support vector machine llamado SVMLeon:

```
SVMLeon <- svm (formula = as.factor (sentimiento) ~ . , data = TrainSparseLeon,
kernel = "linear", scale = FALSE)
```

El primer argumento de la función corresponde a la variable dependiente que es la variable que se quiere predecir, en este caso el sentimiento. Como se trata de una variable dicotómica (-1,1) se le dice a Rstudio que la trate como un factor. Después, se escribe el

símbolo ~ que quiere decir que, seguidamente de la variable sentimiento están todas las demás variables que van a contribuir a explicar la variable dependiente. Las variables que están después de la dependiente son todas las palabras. Para no escribir todas las palabras manualmente, se pone un punto que quiere decir que se usen todas las variables. El segundo argumento indica el nombre del objeto que contiene los datos que se van a usar, que es el subconjunto de datos que corresponde a la parte de entrenamiento. Y por último, se le indica la función de Kernel para la función de decisión que es lineal porque es una línea o un hiperplano el que va a separar las dos clases de la variable dependiente.

Al crear el modelo se han obtenido los parámetros para poder clasificar nuevos tweets, con la función summary se obtiene más información sobre dicho modelo:

```
Summary (SVMLeon)
```

El resultado muestra la siguiente información:

```
Call:

Svm (formula = as.factor (sentimiento) ~ . , data = TrainSparseLeon, kernel = "linear")

Parameters:

SVM-Type: C-classification

SVM- Kernel: linear

Cost: 1

Gamma: 0.01219512

Number of Support Vectors: 43

(30      13)

Number of Classes: 2

Levels:

1        -1
```

Determina que el modelo quiere predecir la variable dependiente "sentimiento" utilizando todas las demás variables. Los parámetros indican que support vector machine es usada

como una máquina de clasificación cuyo núcleo utilizado en el entrenamiento y en la predicción es lineal. El número de vectores de soporte es igual a 43, 30 pertenecen a una clase y 13 a otra. Además, muestra que existen dos clases posibles 1 y -1.

3.3.4 Evaluación del modelo

Una vez que se ha creado el algoritmo de clasificación binaria de support vector machine ya se pueden hacer predicciones. Para ver cómo se comporta el modelo clasificando nuevos tweets, se va a usar el algoritmo creado con el subconjunto de la base de datos correspondiente al entrenamiento, usando los tweets de la parte de evaluación.

Para ello, se escribe la siguiente línea de código donde se ha creado el objeto llamado PredictSVMLeon:

```
PredictSVMLeon <- predict (SVMLeon, newdata = TestSparseLeon)
```

La función anterior tiene como primer argumento el modelo de support vector machine SVMLeon y como segundo argumento los nuevos datos que en este caso van a ser los correspondientes a la parte de evaluación de la base de datos TweetsSparseLeon.

Al cargar esa línea de código, lo que va a hacer Rstudio es realizar las predicciones con el modelo SVMLeon sobre los tweets del subconjunto TestSparseLeon.

3.3.4.1 Matriz de confusión

Para determinar la precisión del modelo se crea una matriz de confusión. Una matriz de confusión es una herramienta estándar para evaluar modelos estadísticos, también se le denomina matriz de clasificación. Su función es ordenar los casos del modelo en categorías, determinando así si el valor de predicción coincide con el valor real (SQL Server, 2016).

Una matriz de confusión tiene la estructura que muestra la Tabla 3.6. En las columnas se encuentran los valores reales, y en las filas los valores de predicción. Se le denomina "verdaderos negativos" cuando el valor real coincide con la predicción y tienen sentimiento negativo. Se le denomina "falsos negativos" cuando el valor real es positivo y la predicción lo ha asignado como negativo. A su vez, los "falsos positivos" son aquellos que en la realidad son positivos y el modelo los ha clasificado como negativos. Los

"verdaderos positivos" son los que en la realidad son positivos y el modelo los ha clasificado como positivos.

Tabla 3.6 Estructura matriz de confusión

Referencia (real) / Predicción	**-1**	**1**
-1	Verdaderos negativos	Falsos negativos (Error Tipo II)
1	Falsos positivos (Error Tipo I)	Verdaderos positivos

Fuente: Elaboración Propia basado en (Zelada, 2017)

En definitiva, los valores que están coloreados en la diagonal son los que el modelo ha clasificado correctamente, los valores que se sitúan fuera de la diagonal son los que el modelo clasificó erróneamente.

Para crear la matriz de confusión en Rstudio se escribe la siguiente línea de código:

```
confusionMatrix (PredictSVMLeon, TestSparseLeon$sentimiento)
```

La función anterior tiene como primer argumento la predicción creada a través del modelo support vector machine, y como segundo argumento se le asignan los valores que en realidad corresponden al sentimiento.

La Tabla 3.7 representa el resultado que proporciona Rstudio de la matriz de confusión. Por un lado, se observa el modelo predijo de manera correcta 32 tweets positivos y 1 tweet negativo. Por otro lado, clasificó incorrectamente 3 tweets clasificados en la realidad como negativos.

Tabla 3.7 Matriz de confusión

Referencia (real) / Predicción	**-1**	**1**
-1	1	0
1	3	32

Fuente: Elaboración Propia a partid de Rstudio

Rsutdio, además, muestra la siguiente información:

```
Accuracy : 0.9167
```

95% CI : (0.7753, 0.9825)

No Information Rate: 0.8889

P-Value [Acc > NIR] : 0.4219

Kappa : 0.3721

Mcnemar's Test P-Value: 0.2482

Sensitivity : 0.95833

Specificity : 0.25000

Pos Pred Value : 0.98462

Neg Pred Value : 0.91429

Prevalence : 0.11111

Detection Rate : 0.027778

Detection Prevalence : 0.027778

Balanced Accuracy : 0.62500

'Positive' Class : 1

La exactitud del modelo según Rstudio es del 91,67%, que se calcula a través de la siguiente formula:

$$Exactitud = (Verdaderos\ Positivos + Verdaderos\ Negativos)/Total.$$

Por tanto, se ha creado un algoritmo que permite clasificar los tweets en positivo o negativo con un poder predictivo del 91,67%.

La sensibilidad hace referencia a la tasa de verdaderos positivos que se calcula a través de la siguiente fórmula:

$$Sensibilidad = Verdaderos\ Positivos/Total\ Positivos$$

El resultado es del 98,46% lo que indica que clasifica los tweets positivos correctamente en ese porcentaje. La misma fórmula se utiliza para determinar la especificidad, es decir, la tasa de verdaderos negativos que muestra que es el 91,42%.

La prevalencia indica la frecuencia en la que un tweet es positivo, en este caso es el 11,11%. En caso de que hubiera el mismo número de tweets positivos que de tweets negativos, la prevalencia sería del 50%.

El índice denominado coeficiente Kappa, es un estadístico que mide la diferencia entre la exactitud que se ha logrado en la clasificación mediante el clasificador automático y la exactitud que se hubiera podido conseguir mediante un clasificador aleatorio (Teledet, s. f.).

Capitulo IV. PROYECTO HUELVA CAPITAL GASTRONÓMICA 2017

Con el objetivo de realizar una comparación entre el sentimiento de la capitalidad gastronómica de la ciudad de León durante el primer semestre del año 2018, se realizará un análisis de sentimiento de la capitalidad gastronómica de la ciudad anterior, Huelva durante el mismo periodo de tiempo, el primer semestre del año 2017.

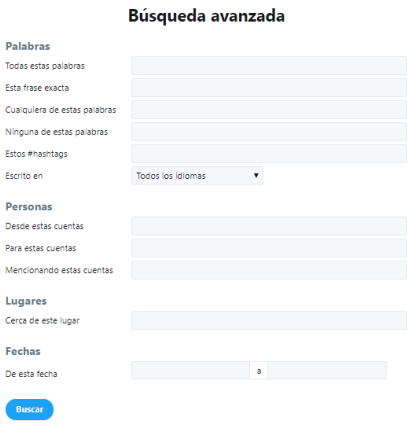
4.1 RECOLECCIÓN DE LOS DATOS

Para la recolección de los tweets, se ha usado la búsqueda avanzada de Twitter ya que la recolección de los datos a través de la nube con Twitter Archiver no es compatible debido a que éste solo recoge los tweets publicados en un periodo de tiempo de 15 días anteriores a partir del momento de la búsqueda. Por ello, resulta imposible acceder a los datos de la capitalidad gastronómica de la ciudad de Huelva durante el primer semestre del año 2017 a través de esta metodología.

La búsqueda avanzada de Twitter tiene el aspecto que se presenta en la Figura 4.1, se observa que permite realizar una búsqueda especifica por palabras, por personas, por lugares y por fechas. Esta aplicación no es automática como sí lo es Twitter Archiver pero permite el acceso a los tweets publicados a lo largo de la historia de Twitter.

Para obtener los datos correspondientes a la capitalidad gastronómica de Huelva se han filtrado los tweets utilizando los hashtags #HuelvaGastronómica2017 y #AhoraCapital.

Figura 4.1 Búsqueda anvanzada de Twitter



Fuente: (Twitter, s. f.)

En base a lo anterior, se ha conseguido una base de datos con un total de 78 observaciones, es decir, 78 tweets que comprendidos durante las fechas de 01/01/2017 al 15/06/2017.

4.2 PREPARACIÓN DE LA BASE DE DATOS

En esta fase se analiza la base de datos con el objetivo de determinar la polaridad de los tweets de una forma manual, es decir, clasificar manualmente el tweet como positivo 1, o negativo -1. El objetivo, como se ha dicho anteriormente, es enseñar al ordenador las claves para que éste sea capaz de comprender los comportamientos y patrones de los datos que hacen que un determinado tweet se clasifique como positivo o negativo, y posteriormente, a través del algoritmo de clasificación sea capaz de predecir el sentimiento de nuevos tweets de una forma automática en pocos segundos.

En definitiva, la base de datos correspondiente a los tweets de la capitalidad de Huelva consta de dos variables, la correspondiente al texto y la correspondiente al sentimiento del tweet.

Una vez preparada la base de datos se transforma a formato csv con separadores por comas para comenzar el análisis.

4.3 PREPROCESAMIENTO DE TEXTO

Como ya se ha dicho, el objetivo de esta fase es la simplificación del texto para que el análisis resulte más sencillo y el aprendizaje del ordenador sea de mayor eficacia.

Con la siguiente instrucción se comprueba cuál es el directorio que utiliza Rstudio para cargar los datos:

```
Getwd()
```

Posteriormente, se crea un objeto denominado TweetsHuelva que será el que contiene los tweets correspondientes:

```
TweetsHuelva <- read.csv ("TweetsHuelva.csv", sep = ";", encoding = "UTF-8")
```

Con el fin de evitar problemas con las tildes y las ñ se escribe el siguiente código:

```
TweetsHuelva$Texto <- chartr ("áéíóúñ", "aeioun", TweetsHuelva$Texto)
```

Una vez que está cargada la base de datos es interesante ver cuántos de los tweets existentes son positivos y cuántos negativos:

```
Table (TweetsHuelva$Sentimiento)
```

El resultado se muestra en la Tabla 4.1:

Tabla 4.1 Recuento sentimiento tweets Huelva

RECUENTO TWEETS HUELVA	
Negativo -1	Positivo 1
12	66

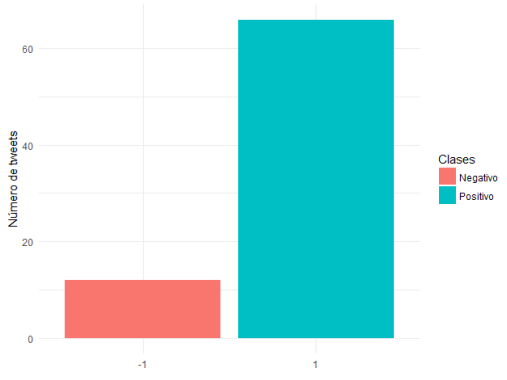
Fuente: Elaboración Propia

Para ver los resultados de una forma más visual se ha generado el siguiente grafico gracias a la librería ggplot2:

```
library("ggplot2")

qplot (data = TweetsHuelva, factor (TweetsHuelva$Sentimiento), geom = "bar",
ylab = "Número de tweets", xlab = "", fill = factor (TweetsHuelva$Sentimiento))
+ theme_minimal() + scale_fill_discrete (name = "Clases", labels = c("Negativo",
"Positivo"))
```

El resultado se muestra en el Grafico 4.1:

Gráfico 4.1 Sentimiento tweets Huelva



Fuente: Elaboración Propia a partir del paquete ggplo2

Antes de comenzar a trabajar con Rstudio hay que asegurarse de que estén cargados todos los paquetes necesarios para la elaboración del análisis de sentimiento. Se genera el corpus con la variable de texto de los tweets de Huelva de la siguiente manera:

```
CorpusHuelva <- Corpus (VectorSource (TweetsHuelva$Texto), readerControl
= list (reader = readPlain, languageEl = "es", load = TRUE))
```

Una vez creado el corpus se comprueba que no tiene errores:

```
Lenght (CorpusHuelva)

[1] 78
```

Efectivamente, es 78 el número de tweets que componen la base de datos.

Ahora, se comprueba que no hay errores dentro del texto de los tweets:

```
Content (CorpusHuelva [50])
```

Se pide que muestre un tweet cualquiera, en este caso se ha optado por el tweet 50, de manera aleatoria, para hacer las comprobaciones. El resultado que muestra Rstudio es el siguiente:

```
[1] "¡Amamos #Huelva y su gastronomía! Aquí tienes los mejores bares y
restaurantes para disfrutar #HuelvaGastronomica17"
```

Efectivamente, el tweet está correctamente con ausencia de errores.

4.3.1 Limpieza general

Se trata de la primera fase del preprocesamiento de texto, es ahora donde hay que hacer una serie de ajustes al corpus.

Se comienza convirtiendo todas las letras de los tweets a minúscula y se hace la correspondiente comprobación:

```
CorpusHuelva <- tm_map (CorpusHuelva, tolower)

Content (CorpusHuelva [50])

[1] "¡amamos #huelva y su gastronomía! aquí tienes los mejores bares y
restaurantes para disfrutar #huelvagastronomica17"
```

A continuación, se eliminan los signos de puntuación:

```
CorpusHuelva <- tm_map (CorpusHuelva, removePunctuation)

Content (CorpusHuelva [50])

[1] "amamos huelva y su gastronomía aquí tienes los mejores bares y restaurantes
para disfrutar huelvagastronomica17"
```

Ahora, se eliminan los números del texto, en caso de que los haya, así como los posibles espacios de más que se hayan generado con la manipulación de los datos:

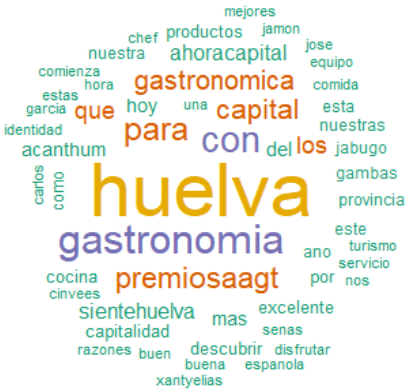
```
CorpusHuelva <- tm_map (CorpusHuelva, removeNumbers)

CorpusHuelva<- tm_map (CorpusHuelva, stripWhitespace)

Content (CorpusHuelva [50])

[1] "amamos huelva y su gastronomia aqui tienes los mejores bares y restaurantes
para disfrutar huelvagastronomica"
```

Una vez concluida la limpieza general del corpus, es interesante crear una nube de palabras para ver cuáles son los términos más frecuentes de dicho corpus:

```
Wordcloud (CorpusHuelva, random.order = FALSE, colors = brewer.pal (6,
"Dark2"), max.words = 80)
```

El resultado se muestra en el Cuadro 4.1 se ve que, igual que en el caso de León, las palabras más usadas son aquellas que han sido utilizadas para la recolección de los tweets, como son "huelva", "gastronomía", "gastronómica", "capital". También hay palabras como "con", "para", "que", "los" que corresponden a las stopwords. En definitiva, son palabras que no aportan significado alguno.

Cuadro 4.1 Nube de palabras CorpusHuelva



Fuente: Elaboración Propia a partir del paquete Wordcloud

4.3.2 Stop word removal

De la misma manera que en capitulo anterior, es el momento de la eliminación de las palabras más frecuentes para la lengua, en este caso el idioma correspondiente es el español:

```
CorpusHuelva <- tm_map (CorpusHuelva, removeWords, c(stopwords ("spanish")))

Content (CorpusHuelva [50])

[1] "amamos huelva gastronomia aqui mejores bares restaurantes disfrutar huelvagastronomica"
```

4.3.3 Steaming

Es la última fase del preprocesamiento de texto y consiste en la reducción de las palabras a su raíz. Para ello, se escribe en Rstudio la siguiente línea de código y se realiza la comprobación:

```
CorpusHuelva <- tm_map (CorpusHuelva, stemDocument, language = "spanish" )

Content (CorpusHuelva [50])

[1] "am huelv gastronomi aqui mejor bar restaur disfrut huelvagastronom"
```

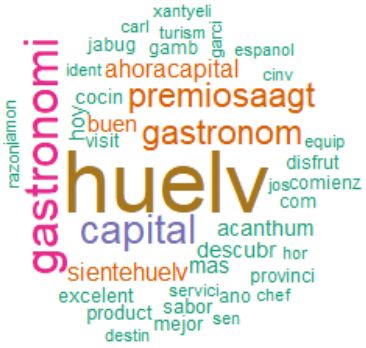
Ahora, se generará otra nube de palabras para ver aquellas con mayor frecuencia una vez reducidas a la raíz:

```
Wordcloud (CorpusHuelva, random.order = FALSE, colors = brewer.pal (6, "Dark2"), max.words = 80)
```

El Cuadro 4.2 muestra el resultado. Se observa que han sido eliminadas las stopwords pero están las palabras que se han usado para la recolección por lo que se procederá a su eliminación para simplificar el corpus y ver así cuales son las palabras más significativas en el mismo. Así mismo se observa que se han reducido todas las palabras a su raíz.

Las palabras que no reflejan sentimiento positivo o negativo y por ello serán eliminadas son: “huelvagastronom”, “huelv”, “capital”, “gastronomi”, “gastronom”, “ahoracapital”, “sientehuelv”, “español”.

Cuadro 4.2 Nube de palabras CorpusHuelva stem



Fuente: Elaboración Propia a partir del paquete Wordcloud

Para la eliminación de dichas palabras se utiliza el siguiente comando:

```
CorpusHuelva <- tm_map (CorpusHuelva, removeWords, c("huelvagastronom", "huelv", "capital","gastronomi", "gastronom", "ahoracapital", "sientehuelv", "espanol"))
```

Para comprobar los cambios de una manera visual se genera otra nube de palabras:

```
Wordcloud (CorpusHuelva, random.order = FALSE, colors = brewer.pal (8,"Dark2"), max.words = 80)
```

El Cuadro 4.3 muestra las palabras más usadas en el corpus una vez realizado el preprocesamiento de texto. Se ve que la palabra más usada es "premiosaagt" seguida por "buen", "mejor", "excelent", "product", "sabor", "cocin", "descubrir", etc.

Cuadro 4.3 Nube de palabras CorpusHuelva stem 2

Fuente: Elaboración Propia a partir del paquete Wordcloud

4.4 ANÁLISIS DE SENTIMIENTO

4.4.1 Creación de la matriz

Antes de comenzar con el análisis de sentimiento propiamente dicho, hay que almacenar los tweets en la base de datos creando una matriz que toma por filas el número total de observaciones, en este caso 78, y por columnas tomará el total de palabras existentes en la base de datos. De esta manera la matriz se rellenará con 0 o 1 dependiendo de si el tweet en cuestión tiene dicha palabra (1) o no (0).

Para ello, se ha creado un objeto denominado FrequenciesHuelva que se trata de una matriz de documento de texto creada a partir del corpus:

```
FrequenciesHuelva <- DocumentTermMatrix (CorpusHuelva)
```

Se inspecciona dicha matriz para obtener más información:

```
FrequenciesHuelva

<< DocumentTermMatrix (documents: 78, terms: 302)>>

Non-/sparse entries: 456/23100

Sparsity: 98%
```

Maximal term length: 18

Weighting: term frequency (tf)

Indica que el objeto creado llamado FrequenciesHuelva es una matriz de documento de texto que consta de 78 documentos y 302 términos, es decir, 302 palabras. Por tanto, es una matriz de 78 filas x 302 columnas. Además, muestra que hay una dispersión del 98% lo que indica que hay gran cantidad de 0 en dicha matriz. También muestra que el largo máximo de un término es de 18, lo que quiere decir que la palabra más grande de la matriz es de 18 caracteres.

Para ver cómo se ha rellenado una parte de la matriz se escribe lo siguiente:

```
Inspect (FrequenciesHuelva [50:55, 55:60])
```

Lo que se pide es que se muestren la fila de la 50 a la 55 y la columna 55 a la 60. El resultado se muestra en la Tabla 4.2 que está completa prácticamente con 0, excepto en la columna 60 correspondiente a la palabra "gamb" cuya fila 52 contiene dicha palabra:

Tabla 4.2 FrequenciesHuelva [50:50, 55:60]

Terms / Docs	cruz	enhorabuen	equip	estren	gabriel	gamb
50	0	0	0	0	0	0
51	0	0	0	0	0	0
52	0	0	0	0	0	1
53	0	0	0	0	0	0
54	0	0	0	0	0	0
55	0	0	0	0	0	0

Fuente: Elaboración Propia

4.4.2 Frecuencia de las palabras

Con el objetivo de determinar cuáles son los términos más frecuentes de la base de datos en cantidades, se transforma el objeto FrequenciesHuelva en una matriz denominada FrequenciesHuelvaMatrix:

```
FrequenciesHuelvaMatrix <- as.matrix (FrequenciesHuelva)
```

Ahora, se suman las columnas para obtener la frecuencia de las palabras:

```
FrequenciesHuelvaMatrix <- colSums (FrequenciesHuelvaMatrix) %>% sort (decreasing = TRUE)
```

Posteriormente, se guardan los resultados en un data frame para su posterior representación gráfica, cuenta con dos columnas; una correspondiente al nombre de las palabras, y otra correspondiente a la frecuencia de las mismas:

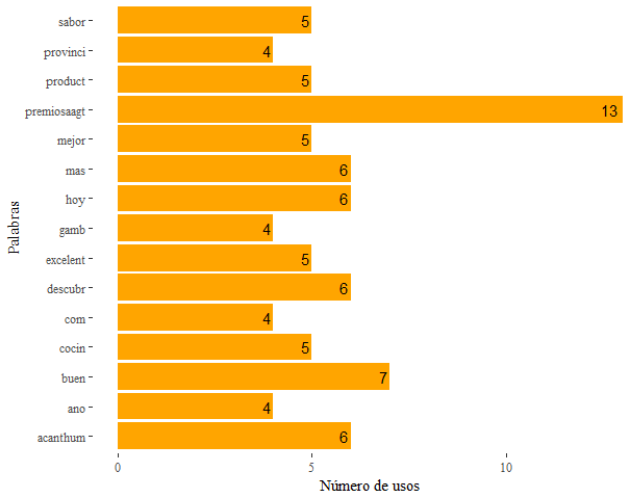
```
FrequenciesHuelvaMatrix <- data.frame (palabra = names (FrequenciesHuelvaMatrix), freq = FrequenciesHuelvaMatrix)
```

Para la representación gráfica se usa la siguiente línea de código:

```
FrequenciesHuelvaMatrix [1:15, ] %>% ggplot (aes (palabra, freq)) + geom_bar (stat = "identity", fill = "Orange") + geom_text (aes (hjust = 1.3, label = freq)) + coord_flip() + labs (,  x = "Palabras", y = "Número de usos") + theme_tufte()
```

Por tanto, las 15 palabras más frecuentes son las representadas en el Gráfico 4.2 Que como ya anunciaba la nube de palabras, la palabra más repetida es “premiosaagt.

Gráfico 4.2 Palabras más frecuentes en Tweets Huelva



Fuente: Elaboración Propia a partir de Rstudio

4.4.3 Reducción de la matriz

Existen palabras en la base de datos que no son relevantes debido a su baja frecuencia, por tanto, dichas palabras no son interesantes para el análisis y hay que suprimirlas. para la reducción de la matriz se ha creado un nuevo objeto denominado SparseHuelva:

```
SparseHuelva <- removeSparseTerms (FrequenciesHuelva, 0.98)
```

Se ha establecido un umbral del 98%, lo que indica que permanecerán aquellas palabras que se repitan en más de un 2% de los documentos. Para obtener más información sobre esta nueva matriz se escribe su nombre:

```
SparseHuelva

<<DocumentTermMatrix (documents: 78, terms: 80)>>

Non-/sparse entries: 234/6006

Sparsity: 96%

Maximal term lenght: 14

Weighting: term frequency (tf)
```

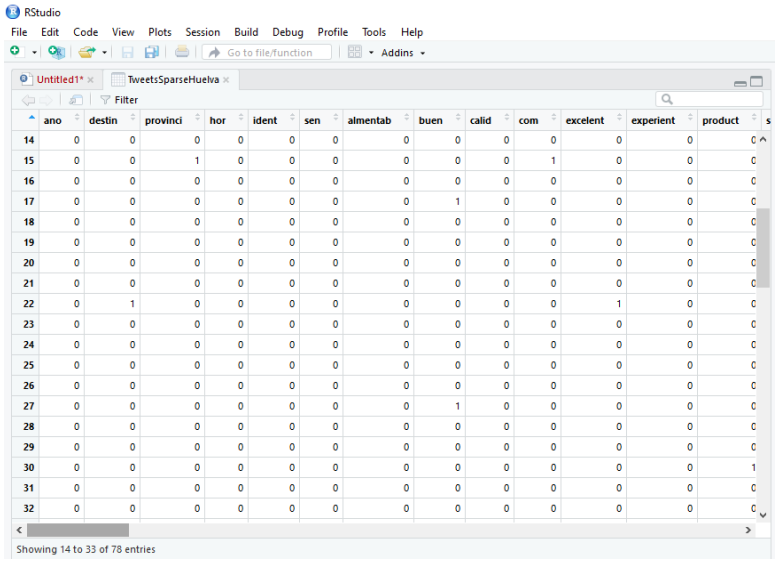
Indica que la matriz creada contiene el mismo número de documentos, 78, pero el número de términos ha disminuido considerablemente a 80. Además, la dispersión se ha reducido a un 96% y el largo máximo de los términos es de 14 caracteres.

Una vez que se ha reducido la matriz, se guardan los resultados en un data frame denominado TweetsSparseHuelva:

```
TweetsSparseHuelva <- as.data.frame (as.matrix (SparseHuelva))
```

La Tabla 4.3 muestra cómo se compone la matriz en formato R.

Tabla 4.3 Data frame TweetsSparseHuelva



	ano	destin	provinci	hor	ident	sen	almentab	buen	calid	com	excelent	experient	product
14	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	1	0	0	0	0	0	0	1	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	1	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	1	0	0	0	0	0	0	0	0	1	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	1	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	1
31	0	0	0	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0

Fuente: Elaboración Propia a partir de Rstudio

En este data frame no existe la variable sentimiento y hay que incorporársela, para ello se asigna dicha variable de la base de datos original TweetsHuelva:

```
TweetsSparseHuelva$Sentimiento <- TweetsHuelva$Sentimiento
```

4.4.4 Creación del modelo de clasificación

Una vez que se ha creado el data frame con la matriz reducida ya es posible la creación del algoritmo de clasificación de support vector machine de machine learning. Como ya se sabe, el objetivo de esta herramienta es la clasificación de los tweets en un lado u otro del hiperplano, clasificándolos como positivos o negativos, según proceda.

Por tanto, se va a usar para dicha clasificación, el margen máximo de clasificación a partir de los vectores de soporte, la misma metodología que se usó en el caso de la capitalidad gastronómica de León.

Se comienza realizando la separación de la base de datos en un 80% correspondiente a la parte de entrenamiento del modelo, y un 20% correspondiente a la parte de evaluación.

Se establece un punto de aleatoriedad para la partición de la base de datos en dos subconjuntos creados aleatoriamente. Para ello se establece la “semilla”:

```
set.seed (12)
```

Una vez definido definido el punto de aleatoriedad, se genera una nueva variable llamada SplitHuelva que decidirá las observaciones que corresponden al conjunto de entrenamiento y aquellas que corresponden al conjunto de evaluación:

```
SplitHuelva <- sample.split (TweetsSparseHuelva$Sentimiento, SplitRatio = 0.8)
```

Al correr el código, Rstudio muestra las observaciones de la siguiente manera:

```
SpliHuelva

[1] TRUE FALSE FASLE TRUE TRUE TRUE TRUE TRUE TRUE …
```

Ahora, se definen los subconjuntos de la base de datos:

```
TrainSparseHuelva <- subset (TweetsSparseHuelva, SplitHuelva == TRUE)

TestSparseHuelva <- subset (TweetsSparseHuelva, SplitHuelva == FALSE)
```

Se comprueba que efectivamente un 80% de las observaciones se han establecido en la parte de entrenamiento y un 20% en la parte de evaluación:

- TrainSparseHuelva tiene un total de 63 observaciones.
- TestSparseHuelva tiene un total de 15 observaciones.

Una vez dividida la base de datos es posible crear el algoritmo de support vector machine.

Se va a crear el modelo de clasificación llamado SVMHuelva:

```
SVMHuelva <- svm (formula = as.factor (Sentimiento) ~., data = TrainSparseHuelva, kernel = "linear", scale = FALSE)
```

Con la creación del modelo ya se han generado los parámetros para la clasificación de nuevos tweets, gracias a la función summary se puede obtener más detalle de este modelo:

```
Summary (SVMHuelva)

Call:

Svm (formula = as.factor (sentimiento) ~ . , data = TrainSparseHuelva, kernel = “linear”,  scale = FALSE)

Parameters:
```

SVM-Type: C-Classification

SVM-Kernel: linear

Cost: 1

Gamma: 0.0125

Number of Support Vectors: 32

(22 10)

Number of Classes: 2

Levels:

1 -1

Muestra que el modelo tiene una variable dependiente, que es la variable a predecir, utilizando todas las demás variables independientes. Además, los parámetros indican que support vector machine es una máquina de clasificación cuyo núcleo utilizado en los dos subconjuntos de la base de datos (entrenamiento y evaluación) es lineal. El número total de vectores de soporte utilizados en el modelo es de 32, 22 pertenecen a la parte positiva y 10 a la parte negativa.

4.4.5 Evaluación del modelo

Una vez que ya ha sido creado el modelo de clasificación de support vector machine utilizando el margen máximo a partir de los vectores de soporte es el momento de hacer predicciones para ver el poder de predicción de este. Para ello, se va a usar la parte de la base de datos correspondiente a la evaluación, TestSparseHuelva, creando el objeto llamado PredicSVMHuelva:

```
PredictSVMHuelva <- predict (SVMHuelva, newdata = TestSparseHuelva)
```

Lo que va a hacer Rstudio es realizar las predicciones utilizando el modelo SVMHuelva pero con los datos de la parte de evaluación.

Por tanto, para conocer la precisión del modelo se genera una matriz de confusión:

```
confusionMatrix (PredictSVMHuelva, TestSparseHuelva$Sentimiento)
```

La matriz de confusión que ha proporcionado Rstudio tiene la forma que muestra la Tabla 4.4, donde se observa que, por un lado, el modelo clasificó correctamente 13 tweets positivos y ninguno negativo. Por otro lado, el modelo clasificó incorrectamente 2 tweets clasificados en la realidad como negativos.

Tabla 4.4 Matriz de confusión

Referencia (real) / Predicción	**-1**	**1**
-1	0	0
1	2	13

Fuente: Elaboración Propia

Rstudio muestra también la siguiente información:

Accuracy : 0.8667

95% CI : (0.5954, 0.9834)

No Information Rate: 0.8667

P-Value [Acc > NIR] : 0.6771

Kappa : 0.3410

Mcnemar's Test P-Value: 0.4795

Sensitivity : 1.0000

Specificity : 0.0000

Pos Pred Value : 0.8667

Neg Pred Value : 0.0100

Prevalence : 0.1333

Detection Rate : 0.0027

Detection Prevalence : 0.0025

Balanced Accuracy : 0.5000

'Positive' Class : 1

La precisión o exactitud del modelo según Rstudio es del 86.67%, lo que indica que clasificará los tweets de forma correcta en positivo o negativo en un 86.67% de los casos.

La sensibilidad hace referencia a la tasa de verdaderos positivos que efectivamente son positivos. Determina que la precisión es del 100% en este aspecto. No obstante, la especificidad hace referencia a la tasa de verdaderos negativos que en este caso marca que es del 0%, porque los tweets negativos que había en el conjunto los clasificó erróneamente.

Capítulo V: ANÁLISIS DE RESULTADOS

En este capítulo se analizarán los resultados del análisis de sentimiento realizado sobre la capitalidad gastronómica de la ciudad de León y del análisis de sentimiento correspondiente a la ciudad de Huelva con el fin de compararlos.

5.1 ANÁLISIS DE LOS MODELOS DE CLASIFICACIÓN

Una vez realizados ambos análisis es interesante y conveniente realizar una comparación, es por ello por lo que en este apartado se van a analizar todos aquellos datos importantes en la realización del análisis utilizando la herramienta de machine learning haciendo una comparación entre León y Huelva.

5.1.1 Análisis datos recolectados

La primera etapa consistía en la recopilación de los tweets, por ello se ha generado una tabla resumen entre las diferencias de ambas ciudades.

En la Tabla 5.1 se observa que, en el caso de León, se han recopilado 178 tweets siendo la mayoría positivos, del mismo modo en el caso de Huelva, aunque el número de tweets recolectados es mucho menor. Además, el porcentaje de tweets positivos es mayor en el caso de León que en el de Huelva, mientras que, el porcentaje de negativos es menor en el caso de León que en el caso de Huelva. Esta información proporciona la evidencia de que la capitalidad gastronómica de León está teniendo mayor impacto y repercusión que la referente a Huelva durante el mismo periodo del año anterior.

Tabla 5.1 Resumen recopilación tweets León y Huelva

RECOPILACIÓN TWEETS	
León (2018)	**(Huelva 2017)**
Nº de tweets recopilados = 178	Nº de tweets recopilados = 78
Tweets positivos = 158 (88,76%)	Tweets positivos = 66 (84,62%)
Tweets negativos = 20 (11,24%)	Tweets negativos = 12 (15,38%)

Fuente: Elaboración Propia

5.1.2 Análisis de la matriz de documentos

La matriz de documentos, como ya se ha dicho anteriormente, está compuesta, en las filas, por el número de tweets, y en las columnas por todas las palabras que hay en todos los tweets. Además, está rellenada por 0 en el caso de que la palabra en cuestión no aparezca en el tweet y 1 en el caso de que sí aparezca. Es por ello interesante analizar la Tabla. Donde se comparan el número de términos y de documentos antes y después de la reducción de dichos términos en las ciudades de León y Huelva.

En la Tabla 5.2 se ve que la matriz de documentos perteneciente a León (FrequenciesLeon) tiene, a priori, 909 términos, mientras que la de Huelva (FrequenciesHuelva) consta de 302, datos lógicos en vista a que el número de documentos es menor en el caso de Huelva. Sin embargo, la dispersión es mayor en el caso de León, lo que indica que hay muchas palabras que no se repiten demasiado. No obstante, una vez reducida la matriz, para el caso de León (SparseLeon) el número de términos se redujo a 82 con una dispersión menor, en el caso de Huelva (SparseHuelva), los términos también se redujeron a 80 con una dispersión del 96%. Esta información indica que León, a priori, tenía más documentos y por tanto más términos, pero una vez que han sido eliminados aquellos menos repetidos, la matriz se ha reducido significativamente. En el caso de Huelva, a priori, no tenía tantos términos como León, pero son más repetidos ya que tras la reducción prácticamente ambos tienen el mismo número de términos.

Tabla 5.2 Resumen matriz de documentos León y Huelva

MATRIZ DE DOCUMENTOS	
FrequenciesLeon (2018)	**FrequenciesHuelva (2017)**
Documentos = 178 ; Términos = 909	Documentos = 78 ; Términos = 302
Dispersión = 99%	Dispersión = 98%
SparseLeon (2018)	**SparseHuelva (2017)**
Documentos = 178 ; Términos = 82	Documentos = 78 ; Términos = 80
Dispersión = 96%	Dispersión = 96%

Fuente: Elaboración Propia

5.2.3 Análisis del algoritmo support vector machine

En base a todo lo anterior, se genera el algoritmo se la máquina de vectores de soporte capaz de clasificar los tweets en un lugar u otro del hiperplano de separación determinando así el sentimiento sobre la capitalidad gastronómica de León y de Huelva.

En la Tabla 5.3 se resume la información obtenida por Rstudio sobre los modelos de clasificación a través de la máquina de vectores de soporte de las ciudades de León y Huelva. Se ve que, en el caso de León el algoritmo de clasificación está más entrenado con 142 documentos que en el caso de Huelva con 63. Esto contribuye a que la precisión del modelo sea más preciso ya que cuantos más ejemplos tenga la maquina mejor serán las predicciones. El número de vectores de soporte obtenidos también es mayor en el caso de León, lo que indica que son más las palabras determinantes a la hora de la clasificación de los tweets en este modelo. Un dato curioso es que el número de vectores de soporte correspondientes a la parte negativa, tanto en León como en Huelva, es prácticamente igual, lo que puede significar que, al no disponer de demasiados ejemplos negativos, la precisión de la predicción de éstos sea reducida.

Tabla 5.3 Resumen Support Vector Machine (SVM) León y Huelva

SUPPORT VECTOR MACHINE (SVM)					
León (2018)			**Huelva (2017)**		
Entrenamiento = 142	Evaluación = 36		Entrenamiento = 63	Evaluación = 15	
	1	-1		1	-1
Nº vectores de soporte = 43	30	13	Nº vectores de soporte = 32	22	10
Precisión = 91,67%			Precisión = 86,67%		
Sensibilidad = 98,46%			Sensibilidad = 100%		
Especificidad = 91,42%			Especificidad = 0%		

Fuente: Elaboración Propia

En cuanto a la precisión del modelo, las evidencias indican que es mayor en el caso de León, y no solo en términos generales, sino que en lo referente a la sensibilidad (tasa de verdaderos positivos) y la especificidad (tasa de verdaderos negativos) dispone de porcentajes de predicción muy elevados. Por el contrario, en el caso de Huelva, la precisión general es buena, pero, como se ha dicho anteriormente, en la parte de

evaluación ningún tweet negativo ha sido clasificado correctamente, por tanto, la sensibilidad es máxima y la especificidad nula.

5.2 ANÁLISIS DE LOS TÉRMINOS MÁS USADOS

Es interesante analizar los términos más usados en cada una de las ciudades gastronómicas para ver qué palabras están asociadas a cada una de ellas.

En la Tabla 5.4 se observa que, antes de realizar el preprocesamiento de texto, las palabras más repetidas en los corpus de León y de Huelva son aquellas que han sido utilizadas para la recolección de los tweets. No obstante, en el caso de Huelva aparece la palabra "premiosaagt" que es muy repetida en los tweets y se refiere a los premios de la Academia Andaluza de Gastronomía y Turismo concedidos a bares y restaurantes de la zona durante el año 2017 (Orden45, 2017).

Tabla 5.4 Resumen de los términos más usados en León y Huelva

TÉRMINOS MÁS USADOS EN EL ANÁLISIS DE SENTIMIENTO	
León (2018)	**Huelva (2017)**
Palabras más usadas en el corpus (antes de modificar): "león", "leonesp", "gastronomía", "manjar", "capital", "leonmanjardereyes"	**Palabras más usadas en el corpus (antes de modificar):** "huelva", "gastronomía", "premiosaagt", "capital"
Palabras más usadas en el corpus (después de la modificación): "mejor", "cecin", "visit", "product", "profesional", "excelent", "cocin", "buen", "promoción", "tap", "delici", "barriohumedo", "gran", "calid"	**Palabras más usadas en el corpus (después de la modificación):** "premiosaagt", "buen", "excelent", "descubr", "mejor", "product", "sabor", "jabug", "gamb", "jamon", "cocin", "chef"

Fuente: Elaboración Propia

En cuanto a las palabras más usadas en el corpus una vez eliminadas las stopwords, reducidas las palabras a su raíz y eliminadas las palabras utilizadas en la recolección, se obtienen unos datos que aportan mayor sentido. Para el caso de León, las evidencias determinan que las palabras más ligadas a la capitalidad gastronómica de esta ciudad tienen que ver con uno de sus productos estrella como es la cecina, además también se repite mucho la palabra cocina, visitar y productos, también se observa el termino tapa,

muy típico de esta zona. Los adjetivos más usados son mejor, excelente, profesional, bueno y delicioso. Para el caso de Huelva, se observa que la palabra más usada es premiosaagt, seguida por productos, sabor, así como el uso de las palabras gambas y jabugo que corresponden a productos representativos de esta zona. Los adjetivos más usados para referirse a la capitalidad gastronómica de Huelva son excelente, bueno y mejor, también presentes en el caso de León.

5.2.1 Análisis tweets positivos

En este apartado se va a realizar un análisis de las palabras que mejor caracterizan a los tweets positivos para determinar las características especiales del sentimiento positivo tanto para el caso de León como para Huelva. Para ello se realizará una nube de palabras en Rstudio y posteriormente una tabla resumen para la comparación de ambas ciudades.

5.2.1.1 Análisis tweets positivos León

Se realizará una nube de palabras para obtener evidencias de una manera gráfica sobre aquellos términos más usados cuando un tweet es positivo. El primer paso es crear un objeto denominado PositivosLeon que se comporta como un subconjunto de la base de datos:

```
PositivosLeon <- subset (TweetsSparseLeon, TweetsSparseLeon$sentimiento ==1)
```

Con la línea de código anterior se ha generado el objeto llamado PositivosLeon que corresponde a un subset de TweetsSparseLeon y que únicamente tiene los tweets con sentimiento positivo, es decir, igual a 1.

A continuación, hay que eliminar la variable "sentimiento" porque en el objeto PositivosLeon únicamente están los tweets positivos, es por ello por lo que no interesa tener la variable "sentimiento":

```
PositivosLeon$sentimiento <- NULL
```

Una vez eliminada la variable, es el momento de hacer cálculos para sacar las frecuencias:

```
PositivosLeon <- as.data.frame (colSums (PositivosLeon)
```

Con este comando se obtendrán las frecuencias creando un data frame compuesto por las sumas de las columnas del objeto PositivosLeon. Se suman las columnas porque en la matriz se ponía un 0 o 1 dependiendo de si la palabra en cuestión aparecía o no en un tweet determinado, entonces están los tweets en filas, las palabras en columnas y 0 y 1 dentro de las casillas. En base a esto, si se suman todos los números de una columna se obtendrá la frecuencia total de esa palabra.

Ahora, hay que definir las palabras como una variable y no como un índice:

```
PositivosLeon$words <- row.names (PositivosLeon)
```

Después de crear la variable anterior, es el momento de renombrar las demás variables:

```
Colnames (PositivosLeon) <- c ("freq", "word")
```

Con ello, ya es posible crear la nube de palabras de los tweets positivos:

```
Wordcloud (PositivosLeon$word, PositivosLeon$freq, random.order = FALSE,
colors = brewer.pal (6, "Dark2"), max.words= 60)
```

El primer argumento de la línea de código anterior es la columna que contiene las palabras, el segundo argumento son las frecuencias. Además, se le indica que muestre un máximo de 60 palabras. El resultado es el siguiente se muestra en el Cuadro 5.1 se ve que la palabra más repetida en los tweets positivos es "mejor" seguida por las palabras "cecin" y "visit" lo que puede indicar que el sentimiento de la capitalidad gastronómica de León en el caso de que este sea positivo está altamente relacionado con la cecina. Además, se observan adjetivos como "buen", "profesional", "excelente", "exquisit" y "delici". Sin duda es un buen indicativo de que la gastronomía leonesa gusta a los que la degustan.

Cuadro 5.1 Nube de palabras tweets positivos León

Fuente: Elaboración Propia a partir del paquete Wordcloud

5.2.1.2 Análisis tweets positivos Huelva

Se repite el proceso del apartado anterior creando una nube de palabras.

Para comenzar, se genera un nuevo objeto llamado PositivosHuelva que corresponde a un subset de la base de datos TweetsSparseHuelva cuya variable de sentimiento es igual a 1:

```
PositivosHuelva <- subset (TweetsSparseHuelva, TweetsSparseHuelva$Sentimiento ==1)
```

A continuación, como ya se han escogido los tweets positivos, se elimina la variable sentimiento:

```
PositivosHuelva$Sentimiento <- NULL
```

Ahora se guardan los resultados en un data frame para hacer los cálculos:

```
PositivosHuelva <- as.data.frame (colSums (PositivosHuelva))
```

Se definen las palabras como nombres y no como índices:

```
PositivosHuelva$words <- row.names (PositivosHuelva)
```

Una vez creada la variable anterior se renombran las variables:

```
colnames (PositivosHuelva) <- c ("freq", "word")
```

Ya es posible generar la nube de palabras:

```
Wordcloud (PositivosHuelva$word, PositivosHuelva$freq, random.order = FALSE, colors = brewer.pal (8, "Dark2"), max.words = 80)
```

El resultado lo muestra el Cuadro 5.2 y se ve que la palabra más usada en el corpus es también la más usada en los tweets positivos “premiosaagt”. Además, están las palabras “excelent”, “buen”, “sabor”, “descubr”. Se ve que en un segundo plano aparecen las palabras “jamon”, “gamb”, “jabug”, “xantyeli”. Lo que indica que el sentimiento positivo sobre la capitalidad gastronómica de Huelva está altamente relacionado con sus productos locales.

Cuadro 5.2 Nube de palabras tweets positivos Huelva

Fuente: Elaboración Propia a partir del paquete Wordcloud

5.2.1.3 Comparación tweets positivos León y Huelva

Con el objetivo de comparar los términos característicos de los tweets positivos de León y Huelva se ha desarrollado la Tabla 5.5 donde se observa que el sentimiento positivo de León está altamente ligado a la cecina, a las tapas, al vino, a la ciudad propiamente dicha. En el caso de Huelva, el sentimiento positivo sobre la capitalidad está altamente ligado a los premios, a las gambas y al jamón de jabugo. En definitiva, en el caso de León se ve que el público también valora como positivo la ciudad en sí misma como un patrimonio especial digno de visitar, aunque en ambas ciudades se valoran altamente los productos locales de la zona.

Tabla 5.5 Resumen análisis tweets positivos León y Huelva

ANÁLISIS TWEETS POSITIVOS	
León (2018)	**Huelva (2017)**
Términos más usados en los tweets positivos: “mejor”, “cecin”, “visit”, “product”, “turism”, “ciud”, “tap”, “winelovers”, “barriohumedo”, etc.	**Términos más usados en los tweets positivos:** “premiosaagt”, “buen”, “excelent”, “sabor”, “product”, “descubr”, “gamb”, “jabug”, “jamon”, etc.

Fuente: Elaboración Propia

5.2.2 Análisis tweets negativos

El objetivo es similar al del análisis de los tweets positivos, solo que en este caso se analizarán aquellos términos que están ligados al sentimiento negativo y cuáles son las

características que mejor lo definen. Para ello, se seguirá la misma metodología que en el apartado anterior.

5.2.2.1 Análisis tweets negativos León

Se repite el proceso:

```
NegativosLeon <- subset (TweetsSparseLeon, TweetsSparseLeon$sentimiento
== -1)
NegativosLeon$sentimiento <- NULL
NegativosLeon <- as.data.frame (colSums (NegativosLeon))
NegativosLeon$words <- row.names (NegativosLeon)
Colnames (NegativosLeon) <- c (“word”, “freq”)
Wordcloud (NegativosLeon$word, NegativosLeon$freq, random.order = FALSE,
colors = brewer.pal (6, “Dark2”))
```

El resultado se muestra en el Cuadro 5.3 donde se ve que no hay tantas palabras como en el caso de los tweets positivos, pero las que hay son bastante reveladoras ya que las más repetidas son “junt” y “promoción” seguidas por “hac” y “hostelería”. Todo ello puede ser un indicativo de que hay personas que tienen un sentimiento negativo de la capitalidad gastronómica de León por quejas de la promoción por parte de la junta y el trato de la hostelería.

Cuadro 5.3 Nube de palabras tweets negativos León

Fuente: Elaboración Propia a partir del paquete Wordcloud

5.2.2.2 Análisis tweets negativos Huelva

Se repetirá el proceso del apartado anterior.

```
NegativosHuelva <- subset (TweetsSparseHuelva,
TweetsSparseHuelva$Sentimiento == -1)
NegativosHuelva$Sentimiento <- NULL
NegativosHuelva <- as.data.frame (colSums (NegativosHuelva))
NegativosHuelva$words <- row.names (NegativosHuelva)
```

```
colnames (NegativosHuelva) <- c ("freq", "word")
```

El resultado de la nube de palabras de los tweets negativos está representado en el Cuadro 5.4, donde se observa que la palabra más repetida en los tweets clasificados como negativos es "francispanieg" seguida de palabras como "cocin", "cociner", "chef", "gamb", "esper", etc. Lo que indica que posiblemente el sentimiento negativo esté relacionado con la cocina de Huelva.

Cuadro 5.4 Nube de palabras tweets negativos Huelva

Fuente: Elaboración Propia a partir de Wordcloud

5.2.2.3 Comparación tweets negativos León y Huelva

Para establecer la comparación entre las ciudades se ha generado la Tabla 5.6 donde se observa que el sentimiento negativo de la ciudad de León está relacionado con quejas de la promoción por parte de la junta y hostelería. Mientras que, en el caso de Huelva, el sentimiento negativo está vinculado con la cocina, por tanto, indica que en León el público que está descontento con la capitalidad es porque no se le otorga la relevancia merecida fuera de León, mientras que en Huelva es por quejas en la cocina.

Tabla 5.6 Resumen análisis tweets negativos León y Huelva

ANÁLISIS TWEETS NEGATIVOS	
León (2018)	**Huelva (2017)**
Términos más usados en los tweets negativos: "promocion", "junt", "hosteleri"	**Términos más usados en los tweets negativos:** "francispanieg", "cocin", "cociner", "chef", "gamb", "esper"

Fuente: Elaboración Propia

CONCLUSIONES

Conclusiones generales

La realización del presente Trabajo de Fin de Grado ha permitido extraer una serie de conclusiones basadas en las evidencias obtenidas durante la elaboración de este.

El uso de los datos para las estrategias de marketing es como el "oro" del que depende el éxito de una empresa u organización, ya que estos datos proporcionan una información que es excepcional y muy valiosa. Además, al vivir en una sociedad hiperconectada a la tecnología la creación de estos datos y su uso es cada vez más frecuente. Lejos están ya las encuestas aburridas y repetitivas pudiendo ser sustituidas por otras herramientas más eficaces como la que se ha desarrollado en este trabajo. Tanto la extracción de la información como la obtención del sentimiento del cliente se pueden automatizar de manera que sea el ordenador el que determine de forma objetiva los resultados. Además, el cliente no es consciente de este análisis por lo que la veracidad de los datos aumenta. Otro de los aspectos muy característicos de este tema es el reducido coste que conlleva el uso de estas técnicas ya que cualquier investigador desde su casa y con un ordenador "normal" puede ser capaz de realizar estudios de mercado muy potentes y fiables.

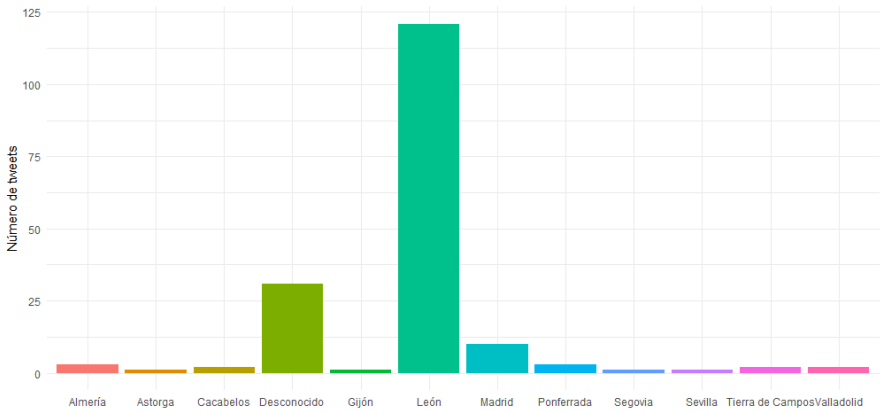
En lo referente a las redes sociales, he concluido que la información que se publica es de carácter público y permanece en el tiempo, es decir, el individuo que en un momento determinado hizo pública una opinión o sentimiento puede no ser consciente de que dicha información tiene un interés empresarial y económico elevado. Muchas son las personas que mencionan la frase "cuando un producto es gratis, el producto eres tú", y razón no les falta ya que en el caso de las redes sociales su descarga es gratuita porque lo realmente valioso es la información que pueden extraer de las personas para fines comerciales. No obstante, es una herramienta de uso común que facilita a sus usuarios expresar sus emociones y sentimientos de una manera libre.

En cuanto al análisis de sentimiento, las evidencias determinan, que se trata de una herramienta muy poderosa que está altamente ligada a la inteligencia artificial lo que tiene un valor añadido ya que se puede automatizar el aprendizaje, y en cuestión de segundos la empresa obtiene el sentimiento del público frente a un determinado producto/servicio, campaña, etc. Además, es una herramienta objetiva ya que la clasificación se hace mediante un algoritmo generado en base a gran cantidad de datos y que, además, es capaz

por si mismo de determinar la polaridad de un texto. Es decir, la inteligencia de negocios hace que los resultados sean más objetivos que si éstos fueran extraídos por personas, objetividad de los algoritmos vs subjetividad de las personas.

En lo referente a la capitalidad gastronómica de León y su correspondiente análisis de sentimiento a través de Twitter, las evidencias determinan que es mucha la oferta de bares y restaurantes ofreciendo sus productos con el argumento de la capitalidad gastronómica, pero no es tanta su demanda por parte del público en general. Además, del conjunto de tweets extraídos la mayoría han sido clasificados como positivos, lo que indica que, aunque no son muchas las opiniones en cuanto a la capitalidad, la mayoría son positivas, por tanto, el sentimiento es en general positivo. Se observan en los gráficos y cuadros que se hace mucha referencia a los productos típicos y característicos de la zona como son la cecina, las tapas o el vino. No obstante, en cuanto a lo negativo muchos son los comentarios que determinan que la capitalidad gastronómica no está teniendo extrema repercusión fuera de León y se quejan del poco esfuerzo de la junta por esta promoción, de ello se hace eco el Grafico 6.1 donde se observa que la gran mayoría de tweets han sido enviados desde León.

Gráfico 6.1 Ubicación Tweets León



Fuente: Elaboración Propia a partir del paquete ggplot2

Haciendo referencia al análisis de sentimiento que se ha realizado para la capitalidad gastronómica del año anterior, Huelva, con el único objetivo de establecer una comparación en términos relativos de la correspondiente a León, las evidencias obtenidas determinan lo siguiente. Las palabras más usadas en este tema tienen que ver con los premios que ha recibido Huelva durante su capitalidad gastronómica, así como la mención

de sus productos más característicos como son las gambas y el jamón de jabugo. El número de tweets recolectados es menor que en el caso de León, pero al igual que éste la gran mayoría son tweets positivos que hacen especial relevancia a sus excelentes productos con gran sabor y a los premios mencionados anteriormente. Los tweets negativos, sin embargo, hacen referencia a todos los aspectos relacionados con la cocina lo que indica que, es probable el sentimiento negativo de los usuarios esté altamente relacionado con quejas en la cocina de los bares y restaurantes de la zona.

En definitiva, tras las evidencias obtenidas, este estudio ha concluido que el sentimiento a través de Twitter sobre la capitalidad gastronómica de León está teniendo cierto éxito en la zona, pero es conveniente otorgar un mayor impulso de esta promoción en las afueras de la región para sacar el mayor partido a este acontecimiento. Esta motivación debería ser promovida por el sector hostelero, pero también por las instituciones y organismos encargados del tema.

Implicaciones empresariales

En base a los resultados del presente trabajo, se observa que muchos son los esfuerzos por parte del colectivo de la hostelería por promocionar sus productos o servicios bajo el contexto de la capitalidad gastronómica a través de Twitter, pero estos esfuerzos no son recíprocos por parte del público. Quizás, los empresarios locales podrían incentivar que sus clientes intercambiaran sus sentimientos promoviendo un hashtag común en sus restaurantes o bares para que éstos expresaran su opinión acerca del producto, del servicio… tratando así de estimular una relación entre el cliente y el empresario más fluída y que éstos sean conscientes de aquellos aspectos que el cliente valora positivamente y de aquellos otros que valora negativamente. Así se optimizarían las estrategias de marketing y se mejorarían los aspectos del branding tanto de bares y restaurantes como de la ciudad en general. Además, el cliente percibe que su opinión importa, y mucho, entonces se sentirá valorado y escuchado. Por su parte, la empresa debería contestar de manera razonada y educada todos los comentarios que surgieran tanto los positivos como los negativos, garantizando así la mejora de la calidad del servicio.

Limitaciones del estudio

El análisis de sentimiento pese a ser una gran herramienta para medir la opinión del público sobre cualquier tema tiene también algunas limitaciones. El lenguaje de las personas es extremadamente complejo ya que existen diferencias culturales, errores gramaticales y ortográficos, el uso de la ironía y el sarcasmo es uno de los principales retos para esta ciencia, así como el contexto puede afectar al tono de expresión de un determinado comentario. Por tanto, que un ordenador sea capaz de comprender todo esto es muy complicado, pero no imposible.

En definitiva, ningún algoritmo tiene un 100% de predicción, como todo, tiene su margen de error y se necesita a las personas para revisarlo.

En lo referente a llevar a cabo el análisis de sentimiento en la práctica, es imprescindible el uso y manejo del software R, lo que ha implicado cierta limitación en el estudio ya que es un programa ciertamente complejo.

Lecciones aprendidas

Tras la elaboración del presente Trabajo de Fin de Grado (TFG) he aprendido la importancia que tienen los datos a la hora de hacer marketing, definitivamente es imposible hacer marketing sin basar las estrategias y esfuerzos en datos. Además, el marketing y la tecnología son grandes aliados para el éxito empresarial o institucional ya que utilizando la inteligencia artificial y el aprendizaje automático se contribuye a construir estrategias de marketing extremadamente valiosas en todos los sentidos.

El análisis de sentimiento utilizando los datos procedentes, en este caso de las redes sociales, permite un conocimiento sobre lo que el público piensa y siente sobre las acciones que lleva a cabo una empresa o institución en tiempo real. Para llevarlo a cabo se necesitan conocimientos de programación, y en este caso de Rstudio, que he adquirido tras una larga fase de preparación antes de comenzar con el trabajo.

Líneas de futuro

La elaboración del presente estudio puede contribuir a la mejora de un marketing de ciudades para extraer un mayor rendimiento a las actividades que pudieran acontecer. Además, puede ayudar a las empresas de diferentes sectores que quisieran dar un nuevo enfoque a las estrategias de marketing utilizando esta herramienta de gran valor.

Hay que recordar que el estudio se ha elaborado con datos correspondientes al primer semestre del año, por tanto, todavía queda otro semestre en el que empresas y ayuntamiento podrían mejorar el sentimiento de la capitalidad gastronómica de la ciudad de León para obtener mayores beneficios económicos y a nivel de marca de ciudad.

REFERENCIAS

Amat Rodrigo, J. (2017). Máquinas de Vector Soporte (Support Vector Machines, SVMs). Recuperado 22 de mayo de 2018, a partir de https://rpubs.com/Joaquin_AR/267926

Berrendero, J. . (2017). Introducción al paquete Caret. Recuperado 18 de mayo de 2018, a partir de https://rpubs.com/joser/caret

Bouchet-Vala, M. (2015). Package SnowballC. *cran.r-project.org*.

Cross Validated. (2018). Número aleatorio Set.seed (N) en R. Recuperado 18 de mayo de 2018, a partir de https://stats.stackexchange.com/questions/86285/random-number-set-seedn-in-r

Fellows, I. (2015). Package «wordcloud».

González, A. (2014). ¿Qué es Machine Learning? Recuperado 9 de mayo de 2018, a partir de http://cleverdata.io/que-es-machine-learning-big-data/

Google Sheets. (s. f.). Twitter Archiver. Recuperado a partir de https://chrome.google.com/webstore/detail/twitter-archiver/pkanpfekacaojdncfgbjadedbggbbphi

Herencia, B. (2018). La era del marketing basado en datos. Recuperado 20 de abril de 2018, a partir de https://blogs.oracle.com/spain/la-era-del-marketing-basado-en-datos

Instituto de ingeniería del conocimiento. (2016). Las 7 V del Big data: Características más importantes. Recuperado 14 de junio de 2018, a partir de http://www.iic.uam.es/innovacion/big-data-caracteristicas-mas-importantes-7-v/

JUrcera. (2012). R y RStudio, instalación y primeros pasos. Recuperado 10 de mayo de 2018, a partir de http://blog.urcera.com/wordpress/?p=242

Kharde, V. A., y Sonawane, S. S. (2016). Sentiment analysis of twitter data: a survey of techniques. *International Journal of Computer Applications*, *139*(11), 975-8887. https://doi.org/10.5120/ijca2016908625

Madroñal Quitín, D. (2015). *Implementación de una Support Vector Machine en RVC – CAL para imágenes hiperespectrales*.

Martínez Gordillo, J. D. (2016). Primer taller de análisis de sentimiento en twitter con R. Recuperado 22 de marzo de 2018, a partir de https://www.youtube.com/watch?v=nOIZnYLlPBo

Max Kuhn Contributions from Jed Wing, A., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., … Max Kuhn, M. (2018). Package «caret». Recuperado 18 de mayo de 2018, a partir de https://cran.r-project.org/web/packages/caret/caret.pdf

Mayer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., y Lin, C.-C. (2017). Package «e1071».

Milton Bache, S., y Wickham, H. (2014). Magrittr: a forward-pipe operator for R. Recuperado 6 de junio de 2018, a partir de https://cran.r-project.org/web/packages/magrittr/index.html

Moreno, A. (2017). Procesamiento del lenguaje natural. Recuperado 18 de abril de 2018, a partir de http://www.vicomtech.org/t4/e11/procesamiento-del-lenguaje-natural

Moreno, G. (2016). El desigual uso de las redes sociales en el mundo. Recuperado 2 de mayo de 2018, a partir de https://es.statista.com/grafico/7325/el-desigual-uso-de-las-redes-sociales-en-el-mundo/

Neuwirth, E. (2015). Package «RColorBrewer».

Orden45. (2017). Andalucía se convierte en el centro de la gastronomía española. Recuperado 6 de julio de 2018, a partir de http://www.orden45.com/premios-aagt16/

Parra, L. (2015). Qué es un csv, cómo se hace y para qué sirve. Recuperado 8 de julio de 2018, a partir de https://lolap.wordpress.com/2015/01/14/que-es-un-csv-como-se-hace-y-para-que-sirve/

Pozzi, F. A., Fersini, E., Messina, E., y Bing, L. (2017). *Sentiment analysis in social networks*. *Sentiment Analysis in Social Networks*. Cambriege (United States). https://doi.org/10.1016/B978-0-12-804412-4.00001-2

R: Data Frames. (s. f.). R: Data Frames. Recuperado 15 de mayo de 2018, a partir de https://stat.ethz.ch/R-manual/R-devel/library/base/html/data.frame.html

Rouse, M. (2017). Inteligencia artificial o AI. Recuperado 1 de julio de 2018, a partir de https://searchdatacenter.techtarget.com/es/definicion/Inteligencia-artificial-o-AI

Rstudio. (s. f.). Data Visualization with ggplot2. *docs.ggplot2.org*, 2.

SAS. (s. f.). *Inteligencia de cliente en la era del marketing basado en datos*.

SQL Server. (2016). Matriz de clasificación. Recuperado 23 de mayo de 2018, a partir de https://msdn.microsoft.com/es-es/library/ms174811(v=sql.120).aspx

Stringfellow, A. (2018). What is data driven marketing? Recuperado 18 de abril de 2018, a partir de https://www.ngdata.com/what-is-data-driven-marketing/

Teledet. (s. f.). Estimación de la exactitud de una clasificación: la matriz de confusión. Recuperado 7 de junio de 2018, a partir de http://www.teledet.com.uy/tutorial-imagenes-satelitales/clasificacion-matriz-confusion-1.htm

Think with Google. (2016). De los datos a la acción: cómo ayuda el marketing basado en datos a la hora de tomar decisiones importantes. Recuperado a partir de https://www.thinkwithgoogle.com/intl/es-es/recursos-y-herramientas/datos-y-metricas/como-ayuda-el-marketing-basado-en-datos-la-hora-de-tomar-decisiones/

tm.r-forge.r-project.org. (s. f.). tm, Paquete minería de textos. Recuperado 10 de mayo de 2018, a partir de http://tm.r-forge.r-project.org/

Tuszynski, J. (s. f.). Package caTools. Recuperado 16 de mayo de 2018, a partir de https://cran.r-project.org/web/packages/caTools/index.html

Twitter. (s. f.). Búsqueda avanzada de Twitter.

Twitter Application Management. (s. f.). Twitter Application Management. Recuperado 8 de julio de 2018, a partir de https://apps.twitter.com/

Vilares, D., Alonso, M. A., y Gómez-Rodríguez, C. (2013). Una aproximación supervisada para la minería de opiniones sobre tuits en español en base a conocimiento lingüístico. *Procesamiento del Lenguaje Natural*, *51*, 127-134.

Wickham, H., y Maintainer,]. (2016). Package «plyr». Recuperado 18 de mayo de 2018, a partir de http://had.co.nz/plyr,

Zelada, C. (2017). Evaluación de modelos de predicción. Recuperado 7 de junio de 2018, a partir de https://rpubs.com/chzelada/275494